ESSAYS ON
SOCIAL NORMS AND ECONOMIC BEHAVIOR

DISSERTATION

submitted to
the Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by

IVO SCHURTENBERGER
from Hochdorf, LU

approved in October 2018 at the request of

Prof. Dr. Ernst Fehr
Prof. Dr. Lorenzo Casaburi

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 24.10.2018
The Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgments

I would like to express my deepest gratitude to my supervisor Ernst Fehr for his continual support, his guidance, and his mentorship. I also thank the dissertation committee member Lorenzo Casaburi for his valuable feedback and constructive questions. Furthermore, I am much obliged to my colleagues at the Department of Economics of the University of Zurich, especially to my coauthors Yagiz Özdemir, Florian Schneider, and Martin Schonger.

I thank my parents Priska and Fabian Schurtenberger for everything they have done for me.

My wife Leandra Schurtenberger deserves special acknowledgment. I am sincerely thankful and humbled to have her in my life.

Table of Contents

Chapter I

An Introduction to Social Norms and Economic Behavior

An Introduction to Social Norms and Economic Behavior

Ivo Schurtenberger

Citation: Schurtenberger, I. (2018). Essays on social norms and economic behavior. *Dissertation*, 6–14.

In our world of conflicting interests, the objectives and desires of two distinct actors are very rarely in perfect accordance. As an illustrative example, consider two firms operating in the same market. Both are interested in selling their product to customers to be profitable. Obviously, the market share of one firm comes at the expense of the other. Moreover, the interests of either firm is not perfectly in line with those of its customers. Customers value high quality and low prices which diminish ceteris paribus the firm's profit. And what about the employees' interests? After all, the firms profit from their employees' effort, commitment and loyalty. This does not mean that their interests must be in perfect alignment, because the success of their firm hinges on the collective effort exerted by all its employees. Therefore, an employee might profit from others' drudging while he himself is simply free-riding along. Also, one employee climbing one step on corporate hierarchy ladder oftentimes means that others are denied from doing so. Additionally, there is not only conflicting interests among employees, but ostensibly also between employee and employer. The former is interested in high wages and agreeable working conditions; requirements that oftentimes cut into the latter's profit. Finally, shareholders may prefer a higher dividend over the firm's management's aspiration to reside in a new and prestigious headquarters in the best district of the city. This example can be readily extended to very different actors, for instance, two countries negotiating a trade deal, two armies facing one another on the battlefield, a paleolithic tribe hunting large game or raising children, and husband and wife deciding whether to watch football or go to the opera.

What would this world of conflicting interests look like if Homo economicus would inhabit it? Homo economicus, the perfectly rational and narrowly self-interested archetype in economics, neglects society when making decisions and has no moral compass guiding his behavior whatsoever. This model serves economics well in many areas due to its rigor, simplicity, and the precision of its predictions. Self-interest is indisputably *one* of the fundamental drivers of human behavior. Notwithstanding, a world inhabited by Homo economicus is most likely best described as *bellum omnium contra omnes*—the constant state of war (or struggle) of each against all (Hobbes, 2005 Orig. pub. 1651). It would truly resemble a dog-eat-dog world. One crucial aspect is worth stating explicitly at this point; Homo economicus does neither care about the positive nor the negative impact his actions have on his fellow men, that is, he would not even incur minimal costs for someone else's dramatic benefit nor would he refrain from an action that benefits himself slightly but would constitute the demise of the counterpart.

Such a world would clearly lack the striking success humankind has achieved. This prosperity is in a large part due to our ability to cooperate in social dilemmas, that

is, situation where taking the strictly self-interested dominant strategy leads to socially undesirable and ultimately inefficient outcomes. We encounter mundane examples of cooperative actions every single day for instance when collaborating with our co-workers. Our ancestors cooperated to hunt large game, raise children, or wage war against other tribes. Cooperation on such massive scale is unique to Homo sapiens. It enabled humans to walk a celestial object different from the one they evolved on. It has brought forth great inventions such as reinforced concrete, the automobile, vaccination, the jet engine or the Internet. It facilitated the rise of nations, which in turn may reinforce cooperation, enabled the formation of organizations and sustained the spreading of fundamental ideas such as the human rights or democracy. All of this stems from the fact that Homo sapiens—the wise man—dwells the earth instead of Homo economicus—the economic man.

Would Homo economicus not come up with a set of institutions that holds his and his fellows' selfish drive in check such that all can prosper? A brief look at human history lets one doubt that the most important institutions in place nowadays could have ever emerged without the massive degree of selflessness—even self-sacrifice—that their establishment and/or continuity required time and again. For instance, would Homo economicus stand in line with the Athenian hoplites in Marathon to defend the young democracy which would later greatly shape Europe and the Western world? Would he step in front of a tank rolling down Tiananmen Square? Would he join hands with about 2 million people to form the Baltic Chain to demonstrate a desire for freedom and independence from Soviet oppression? There are myriad examples of great men and women who drafted constitutions and laws, who defended freedom and democracy against tyranny and oppression, who founded or worked for humanitarian organizations. Their actions usually cannot be explained by pure self-interest.

Homo sapiens differs fundamentally and systematically from Homo economicus in crucial domains. Humans exhibit extraordinary sociality and differentiating right from wrong—even more profoundly: good from evil—has long been a key interest of the species. Therefore, it should not be surprising that "norms" are one of the most invoked concepts in the social sciences (Fehr & Fischbacher, 2004b). These *commonly known standards of behavior that are based on widely shared views of how individual group members ought to behave in a given situation* (Elster, 1989a; Fehr & Fischbacher, 2004a; Bicchieri, 2006) have been argued to play a key role in several economically relevant domains (Akerlof, 1980, 2007; Cialdini et al., 1991; Bernheim, 1994; Conlin et al., 2003; Krupka & Weber, 2013); amongst them the domain of social dilemmas or externalities more generally (Ostrom, 1998).

This dissertation encompasses four individual research projects that examine economic behavior and social norms. Chapter II reviews the literature on the normative foundation of human cooperation. Chapter III studies the enforcement of social cooperation norms under imperfect information. Chapter IV examines the dynamics of norm formation and norm decay as well as the causal effect of social norms on behavior in a collective action context. Finally, chapter V investigates immoral labor markets, employment that produce negative impact, and the role of moral persuasion and dissuasion in it.

In more detail, chapter II reviews the hypothesis that social norms shape human cooperation. First, we identify ten fundamental regularities found in cooperation experiments. We then pose the question whether social norms can accommodate these behavioral patterns and assert that a norm of conditional cooperation—the prescription to at least match others' level of cooperativeness—is indeed consistent with many of the fundamental regularities. However, we emphasize that "ad hoc" social norms can almost always explain behavior, which renders this approach void. In other words: what we refer to as the direct social norm approach requires discipline in form of identification of social norms. We show that several methods for eliciting social norms exist and that they in fact reveal a conditional cooperation norm. Furthermore, we discuss the relationship between the direct social norm approach and theories of social preferences. Fundamentally, we put forward the notion that social preferences are individuals' intrinsic motives, for instance, fairness or reciprocity guiding behavior, but it is the social norm, a collective entity, that defines what is perceived as fair or kind in a given situation. Hence, social preferences are decisive for norm compliance as well as the willingness to sanction free-riders. We present evidence that social norms are often causal drivers of human cooperative behavior and that normative constraints play a crucial role in enforcement behavior and its efficacy. Finally, experiments that allow subjects to vote by foot reveal a preference for institutions that allow for sanctions and the normative guidance of cooperation and constraints on enforcement.

In Chapter III we examine the enforcement of social norms of cooperation under imperfect monitoring. The collapse of cooperation over time in social dilemmas is a well-known and frequently replicated finding if factors such as communication and enforcement are experimentally ruled out (e.g. Isaac et al., 1985; Kim & Walker, 1984; Andreoni, 1995; Ostrom et al., 1992). Fehr & Gächter (2000a) show that costly peer punishment prevents this gradual breakdown of cooperation. Nevertheless, peer punishment exhibits several problems such as anti-social punishment (e.g. Cinyabuguma et al., 2006; Herrmann et al., 2008; Gächter & Herrmann, 2009, 2011), free-riding on altruistic punishment of others (Fehr & Gächter, 2002) and counter-punishment (e.g. Nikiforakis, 2008; Nikiforakis &

Engelmann, 2011). A centralized authority may not be prone to such problems (Baldassarri & Grossman, 2011). Another argument brought fourth is that the laboratory experiments about peer-punishment and cooperation are conducted under the unrealistic assumption of perfect information. Several studies relax this assumption and indeed find that imperfect information about the actions of group members may pose problems for peer-punishment to resolve social dilemmas (e.g. Ambrus & Greiner, 2012, 2015; Grechenig et al., 2010; Fischer et al., 2013; Nicklisch et al., 2015).

We experimentally compare the efficacy of a decentralized peer punishment institution with a centralized one in a public goods setting under various imperfect information environments, that is, subjects may mistake fellow cooperators as defectors and vice versa. In some treatments we further provide subjects with the possibility to costly acquire new information about the actual actions of their group members. We find that decentralization outperforms centralization when information is imperfect and private, that is, every group member receives an individual private signal about others' action, and subjects have the opportunity to buy further signals. Subjects make ample use of improving the information base by costly acquiring new signals. Many subjects buy new signals when their group members are depicted as "defectors," but very few when the signal states "cooperator." This leads to less wrongful punishment of the innocent and to higher cooperation rates compared to an environment where the gathering of more information is experimentally ruled out. We find that the trade-off between centralization and decentralization is characterized by a pattern of frequent wrongful punishment of the innocent (Type-I error) but infrequent acquittal of defectors (Type-II error) under decentralized peer punishment, whereas relatively speaking the opposite holds true for the centralized institution. This superiority is not driven by the fact that aggregate information of those who hold the power to sanction is better under decentralization when signals are private. On the contrary, private signals as opposed to public signals are rather a curse than a blessing for decentralized institutions; this distinction does not affect the performance of the centralized authority. Taken together, decentralization is superior to centralization in the enforcement of social norms of cooperation under all three information structures considered: exogenous imperfect public information, exogenous imperfect private information, and endogenous imperfect information.

Chapter IV studies the dynamics of norm formation and norm decay and their causal influence on economic behavior in social dilemmas. We use a single experimental feature to address both questions in a laboratory public goods experiment. Specifically, we provide subjects, in some treatments, with the opportunity to form a social norm about what constitutes appropriate contributions to a public good by asking them how much

each group member should contribute to the common cause according to their opinion. The mean of subjects' answers is conveyed to the whole group.

We observe that regardless of the presence of peer punishment opportunities subjects on average believe that group members should contribute a substantial fraction, close to the surplus maximizing level, of their endowment to the public good. Over time, when enforcement is possible, group members quickly reach a consensus about how much should be contributed. They do not only agree that group members should contribute close to the maximal extent, but they also obey this prescription. In stark contrast, without norm enforcement in form of peer punishment, disagreement about appropriate behavior spreads and compliance with the average request strongly crumbles over time. Hence, a strong and stable consensus demanding high contributions emerges quickly under peer punishment, which is subsequently largely obeyed. Disagreement, on average lower requests, and strong disobedience characterize the groups that have no access to peer sanctioning. We run a second set of experiments that explicitly elicit the social norms and their change over time in our setting. This empirical assessment clearly reveals that in the beginning the social norm is to contribute the or close to the surplus maximizing level. Consistent and obeyed normative requests demanding high contributions let the social norm further solidify such that after continued interaction high contributions become more appropriate and medium and low contributions less so. In striking contrast, declining and violated normative requests—as is regularly the case if there are no means of enforcement—ultimately cause the social norm to wither. This makes medium contributions the most desirable behavior, even more so than high contributions, and low contributions much less inappropriate.

Regarding the second question, there is a positive causal effect of the norm formation opportunity on cooperation rates, but only for groups with a sanctioning system. We do not find a positive impact of the norm formation opportunity for those groups that lack the sanctioning institution. Not only does the norm formation deceive increase contributions but the groups also benefit with regard to efficiency, that is, groups with the device earn on average more. Moreover, the norm formation opportunity renders peer punishment more efficient in terms of earnings than no punishment, which is not the case if this opportunity is absent.

Finally, in chapter V we explore moral persuasion and dissuasion in immoral labor markets. Several industries, such as arms trade or tobacco, are commonly regarded as immoral (Frank, 1996; Brun et al., 2017). Making one's living as an employee of a corporation operating in such an industry may therefore cause substantial self- and social-image costs

for which an employee must be compensated. This may partly drive that corporations and their adversaries exert considerable effort to mange the perception of said industries (Szczypka et al., 2007; Müller & Kräussel, 2011; Kotchen & Moon, 2012).

We use a laboratory experiment to study whether popular persuasion and dissuasion attempts are effective in changing labor supply for jobs that are perceived as immoral. In the laboratory, the job consists of wrapping three cigarettes into gift wrap paper and placing them into a gift bag which is distributed to young adults resembling freebies at a marketing event. We elicit subjects' reservation wage for accepting this job in an incentive compatible way. Depending on the treatment subjects are either exposed to the company video of the cigarette manufacturer, an anti-tobacco video produced by a large NGO or a neutral control video. The company video highlights the company's role in providing its workforce with a purpose in life, its responsible business practices and how it improves the living conditions in the communities it operates in. The anti-tobacco video stresses the severe health consequences including the annual global death toll and the industry's attempts to addict new (young) customers. Social-image concerns are incorporated by showing subjects' portraits to everyone else in the experimental session next to their decision whether or not to accept the immoral job.

The data reveals the following results. First of all, we find tremendous heterogeneity in reservation wages. About a quarter of subjects accept the job for CHF 1 or less, about another quarter of subjects refuse this job even for the maximal wage of CHF 25. Second, individuals' own normative views and their beliefs about the social appropriateness of accepting the job can explain reservation wages. Third, the company video does not shift subjects' own normative judgment about working for the tobacco manufacturer nor how socially appropriate they regard accepting the job in lab is seen. Subsequently, we do not find a significant difference in the labor supply between those subjects who have watched the company video instead of the neutral control video. Fourth, the dissuasion effort significantly lowers the social appropriateness of accepting the job, although not substantially enough to shift in the labor supply significantly.

This dissertation contributes to the economic literature by expanding our understanding of human behavior—the way Homo sapien behaves—and how it systematically differs from the one of Homo economicus. This helps building a more elaborate economic theory providing us with more accurate predictions about economic behavior. These four independent research papers are therefore deeply rooted in the tradition of behavioral economics. More generally, the insights gained from the presented research strengthens and refines the view that human behavior rests on a normative foundation intertwined

with societal drivers. Core economic activities, namely collective action and labor market participation, have been demonstrated to be closely connected to social norms. Humans care about the impact—negative and positive—their actions exhibit. Social norms play a fundamental role in ensuring that conflicting interests are managed in a civilized and sophisticated manner rendering life as we know it possible. Their utmost importance might be fathomed in the juxtaposition of contemporary life and the horrors roaming earth's surface when basic social norms collapse as they may in the face of civil war or the disintegration of a nation into anarchy.

Chapter II

Normative Foundations of Human Cooperation

Normative Foundations of Human Cooperation

Ernst Fehr & Ivo Schurtenberger

Abstract

A large literature shares the view that social norms shape human cooperation, but without a clean empirical identification of the relevant norms almost every behaviour can be rationalized as norm driven, thus rendering norms useless as an explanatory construct. This raises the question of whether social norms are indeed causal drivers of behaviour and can convincingly explain major cooperation-related regularities. Here, we show that the norm of conditional cooperation provides such an explanation, that powerful methods for its empirical identification exist and that social norms have causal effects. Norm compliance rests on fundamental human motives ('social preferences') that also imply a willingness to punish free-riders, but normative constraints on peer punishment are important for its effectiveness and welfare properties. If given the chance, a large majority of people favour the imposition of such constraints through the migration to institutional environments that enable the normative guidance of cooperation and norm enforcement behaviours.

Citation: Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7), 458.

Normative constraints and prescriptions are ubiquitous and pervade almost every aspect of human social life, from the mundane to the most profound. They appear to play a role in all social groups and have been documented for a large number of ancient societies (Boyd & Richerson, 1994; Sober & Wilson, 1999), but also play a role in contemporary societies. Norms are part of the weave of social life and, if obeyed, they make it predictable, constitute social order and become the cement of society (Elster, 1989a), but if compliance with fundamental norms breaks down—as it sometimes happens in the aftermath of lost wars or natural disasters—disorder, revolt or revolutionary chaos prevails, and life becomes "solitary, poor, nasty, brutish and short" (Hobbes, 2005 Orig. pub. 1651).

Human cooperation is an equally ubiquitous phenomenon that is present in some form in almost every social relationship and is key for the success of social units from the family to the nation state to global organizations (Fehr & Fischbacher, 2003). Sometimes, cooperation is in the material self-interest of people, but here we are interested in those aspects of cooperation where economic incentives alone are not sufficient to induce individuals to cooperate because free-riding would maximize their private gains. Throughout human history, myriad scenarios are characterized by such social dilemmas. Every successful sequential exchange, in which one party provides the quid pro quo first, constitutes an act of cooperation. Our ancestors also faced social dilemmas when they hunted large game, during tribal warfare or during reciprocal food sharing in times of need. Contemporary humans encounter them in team production settings and whenever there is a tension between one's own interest and the reputation of the company, when paying taxes despite low probabilities of being caught in tax evasion or in the context of problems of a truly global scale such as climate change.

To what extent and how do social norms shape human cooperation? There are social norms, such as the norm to keep a promise or the honesty norm, that affect behaviour in cooperative contexts, but are not directly related to cooperation. For example, the honesty norm proscribes lying and that implies that one should also not lie to evade taxes and the norm to keep one's promises implies that one should also keep promises made to an exchange partner, but these norms have implications that go far beyond cooperative contexts. In this Review, we focus instead on social norms that directly prescribe, and limit their prescription to, cooperation and punishment behaviours in social dilemma and collective action contexts. An example of such a norm is the 'conditional cooperation norm', which we define in more detail below. We ask whether these norms can, in principle, explain major behavioural regularities observed in collective action contexts, what the properties of these norms are and which motivational forces ensure compliance

with them, and whether they indeed guide or are the causal drivers of behaviour in collective action.

To answer these questions requires a clear definition of social norms. We define them as *commonly known standards of behaviour that are based on widely shared views of how individual group members ought to behave in a given situation* (Elster, 1989a; Fehr & Fischbacher, 2004a; Bicchieri, 2006). This definition entails three crucial features of social norms. First, a social norm establishes a normative standard of behaviour that applies to a particular group and to a particular situation. Second, the norm is not defined in terms of group members' actual behaviour nor in terms of their motives, their compliance or the conditions under which compliance occurs; it is exclusively defined in terms of a normative behavioural standard, that is, how group members ought to behave. Third, this normative standard and its widely shared approval is commonly known by group members.

Because a norm requires that the normative standard is widely shared, non-compliance with the norm automatically triggers some disapproval. Therefore, if individuals dislike the thought that others disapprove of them, they automatically have some incentive to comply, although, as we will see, this incentive may not necessarily be sufficient to induce compliance. We will therefore also ask which kind of other motives and mechanisms support compliance with social cooperation norms and whether they act as a constraint on potential non-compliers or are part of the 'intrinsic' motivation of individuals. In this context, we will also ask whether the (peer) punishment of norm violators is itself a social norm or whether it is driven by other motivational sources.

1 Regularities in cooperation-related behaviours?

To assess the role of social norms for human cooperation, we first describe major behavioural regularities observed in experimental social dilemma games. With the exception of experiments that allow for face-to-face communication, the subjects in these games are anonymous to each other. They play for real money under conditions where complete free-riding is the dominant strategy for selfish individuals in one-shot games and backward induction implies that complete free-riding is also predicted in the finitely repeated game. We deliberately restrict ourselves to these experimental settings because to precisely identify the role of social norms, their predictions must differ from the self-interest model. Field evidence, in contrast, typically does not allow self-interest to be ruled out with perfect certainty, but below we point out that many lab observations resemble reg-

ularities that are observed in naturally occurring environments. Second, we discuss the ability of social norms to provide a parsimonious explanation for the regularities.

The following patterns are among the key findings in the literature:

1. Although complete free-riding is a dominant strategy, a substantial share of the subjects cooperate in one-shot social dilemmas but free-riding frequently also prevails (Dawes et al., 1977; Dawes, 1980). However, if subjects can communicate about the game before they play it cooperation strongly increases (Dawes et al., 1977; Isaac & Walker, 1988; Sally, 1995) (Fig. 1 a).

2. A large proportion of subjects are conditional cooperators, that is, the belief that other group members cooperate at high levels induces them to also cooperate at high levels but if others are believed to decrease their cooperation these individuals also decrease their cooperation (Fischbacher et al., 2001; Kocher et al., 2008; Chaudhuri, 2011) (Fig. 1 b).

3. In finitely repeated public good games (PGG), cooperation is initially relatively high but often declines to very low levels towards the final periods (Isaac et al., 1985; Kim & Walker, 1984). This holds regardless of whether the game is frame as a public goods game or as a common pool resource game (Andreoni, 1995). If subjects play the finitely repeated game several times—but each time with a new composition of group members—cooperation always starts high and becomes very low towards the end of the game (Ambrus & Pathak, 2011) (Fig. 1 c).

4. In finitely repeated public good games, cooperation is generally higher in groups with a stable group composition ("partner matching") compared to random reassignment of individuals to groups in every period ("stranger matching") (Chaudhuri, 2011; Ambrus & Pathak, 2011; Croson, 1996) (Fig. 1 d).

5. Merely framing a simultaneously played prisoners' dilemma game differently by calling it Community Game instead of Stock Market Game typically causes substantial increases in cooperation rates. However, if the game is played sequentially this framing effect vanishes (Ellingsen et al., 2012; Liberman et al., 2004) (Fig 1 e).

6. There is a widespread willingness to punish free-riders even in one-shot interactions although it is costly for the punisher (Fehr & Gächter, 2000a, 2002; Fehr & Fischbacher, 2004b) (Fig. 2 f). Furthermore, peer punishment opportunities in repeated interactions cause large cooperation increases and often lead to near complete and stable cooperation under partner matching (Fehr & Gächter, 2000a;

Gächter et al., 2008) (Fig. 2 d). These opportunities are, however, also associated with high initial costs such that group welfare does not increase (or even decreases) for roughly 10 periods (Fehr & Gächter, 2000a; Gächter et al., 2008).

7. The effectiveness of peer punishment in enhancing cooperation is undermined if punishment threats signals selfish intentions (Fehr & Rockenbach, 2003; Fehr & List, 2004; Houser et al., 2008; Xiao, 2013) and by "perverse" (Cinyabuguma et al., 2006) or "antisocial" punishment (Herrmann et al., 2008; Nikiforakis, 2008) of cooperators in public good games by those who free-ride—a tendency that varies strongly across different cultures (Fig. 2 b).

8. Despite the high initial cost caused by peer-punishment, subjects eventually prefer environments with a peer punishment opportunity almost unanimously over an environment that rules out peer punishment (Gürerk et al., 2006, 2014) (Fig. 2 c).

9. The opportunity to reward cooperators—either through the preferred choices of cooperative partners (Brown et al., 2004) or through the direct rewarding of those with a high reputation for cooperation (Rockenbach & Milinski, 2006; Sefton et al., 2007; Rand et al., 2009; Balliet et al., 2011) causes large cooperation increases (Fig. 2 d).

10. Stable cooperation at very high levels can be achieved when (i) cooperative individuals are exogenously matched together (Gächter & Thöni, 2005; Kimbrough & Vostroknutov, 2016) (Fig. 2 e) or (ii) in intergenerational public good games when individuals can give advice that is common knowledge to the next generation (Chaudhuri et al., 2006).

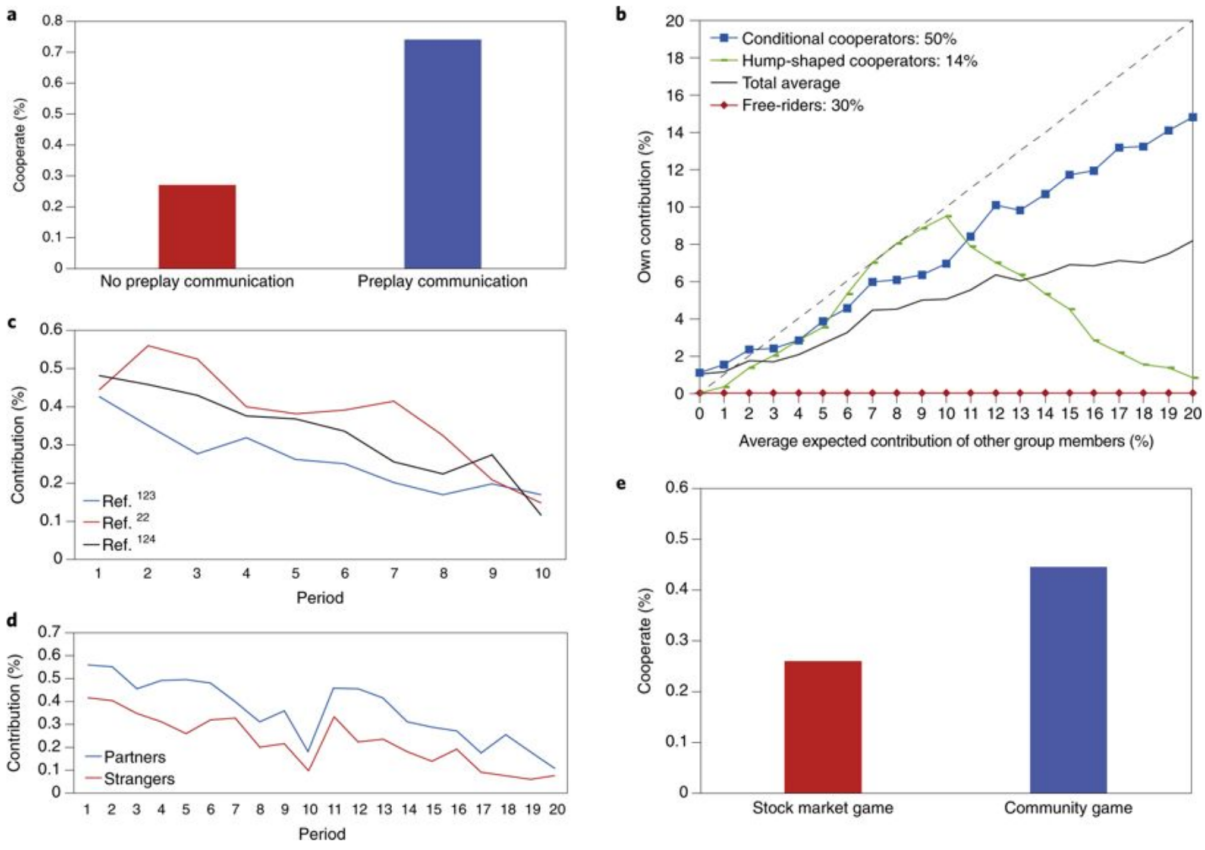


Figure 1: Illustrations of behavioural regularities 1–5 in cooperation experiments.

Fig. 1. **a**, Cooperation rates in a one-shot social dilemma game with and without pre-play communication among the subjects (regularity 1; Dawes et al. (1977)). **b**, Higher expectations of other group members' cooperation causes on average an increase in individual's own cooperation (regularity 2), but individuals are heterogeneous with, typically, a majority of conditional cooperators, a significant minority of full free-riders and some share of hump-shaped conditional cooperators (Fischbacher et al., 2001). The dashed line is the 45° line. **c**, Decline in cooperation rates over time in finitely repeated public goods experiments in which free-riding is the payoff maximizing strategy for selfish subjects (regularity 3; Fehr & Gächter (2000a); Isaac & Walker (1984); Andreoni (1988)). **d**, Cooperation rates in partner treatments are typically higher than those in stranger treatments. In this study, subjects initially believed they had to interact for ten periods after which the experimenter implemented a surprise restart of the same ten-period experiment (regularities 3 and 4; Croson (1996)). **e**, Merely calling the prisoners' dilemma a community game—as opposed to a stock market game—increases cooperation (regularity 5; Ellingsen et al. (2012)) but if the game is played sequentially, this framing effect vanishes.

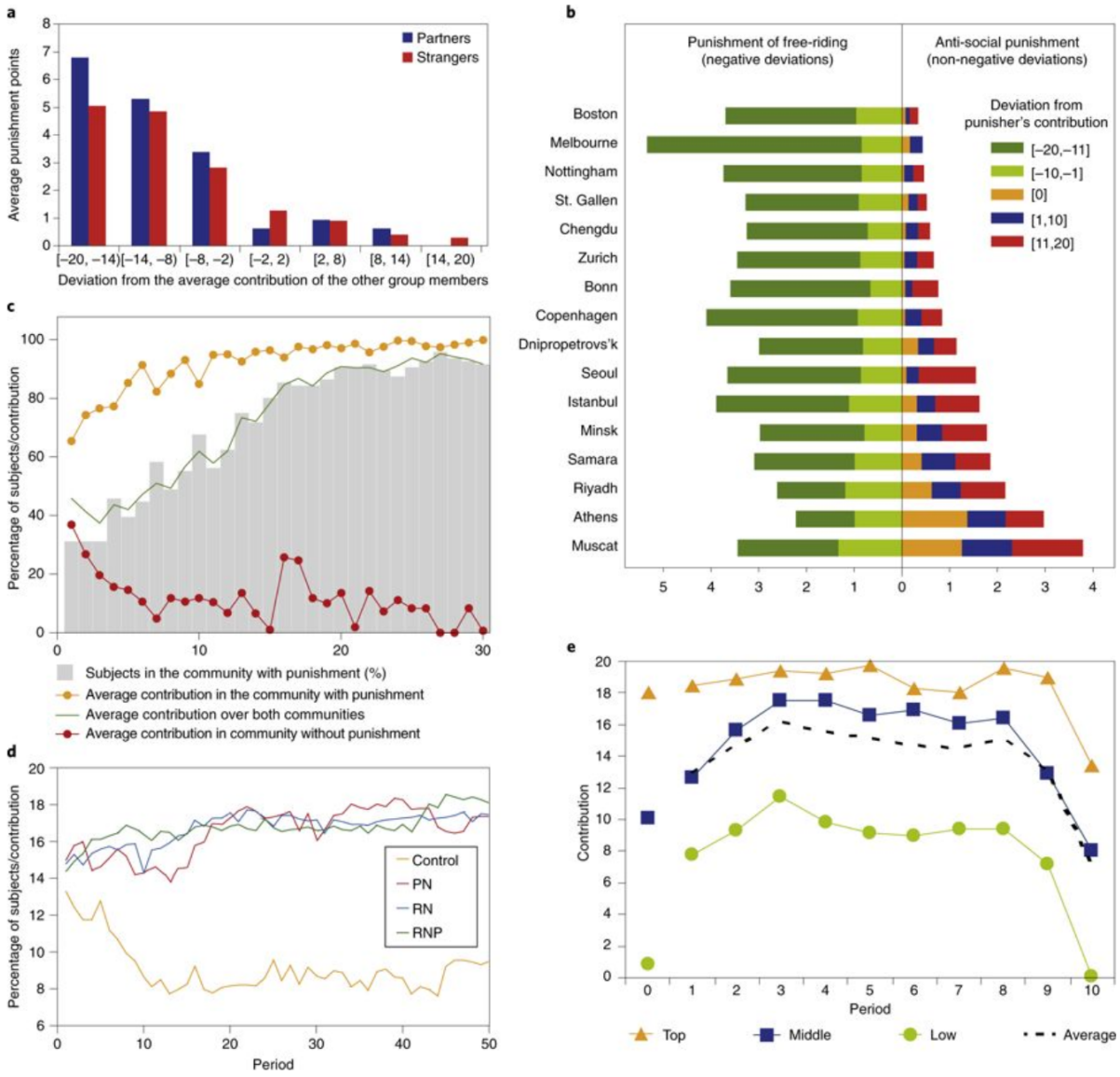


Figure 2: Illustrations of behavioural regularities 6–10 in cooperation experiments.

Fig. 2. **a**, Punishment of group members—measured in terms of the experienced percentage reduction in income—as a function of the deviation of their cooperation level from the average cooperation of other group members (regularities 6 and 7; Fehr & Gächter (2000a)). Punishment of free-riders is very high but above average cooperators also face some 'perverse' punishment. **b**, Evidence for strong cultural differences in antisocial punishment of cooperators (regularity 7; Herrmann et al. (2008)). **c**, Subjects could choose in every period whether they want to be in the community with a peer punishment opportunity or the community without this opportunity. The vast majority of subjects eventually preferred the community with peer punishment (regularity 8; Gürerk et al. (2014)). **d**, The opportunity to punish peers after they observed others' cooperation levels (treatment PN) leads to large increases in cooperation relative to a control treatment without peer punishment (control). The opportunity to mutually reward each other (RN) leads to similarly high cooperation levels compared with PN and treatments with both reward and punishment (RNP) (regularity 6 and 9; Rand et al. (2009)). *e*, High cooperators in a one-shot prisoner's dilemma are grouped together in a subsequent ten-period public goods game. Likewise, the middle and the low cooperators are grouped together. High cooperators achieve very high cooperation rates during the first nine periods (regularity 10; Gächter & Thöni (2005)).

An important question is how insights gained in lab experiments relate to behaviour in naturally occurring environments. Several studies (Barr et al., 2014; Fehr & Leibbrandt, 2011; Rustagi et al., 2010; Keizer et al., 2008; Kosfeld & Rustagi, 2015; Breza et al., 2018; Kaur, forthcoming; Gelcich et al., 2013; Burks et al., 2016; Carlsson et al., 2014) demonstrate that individuals' behaviour in the lab is predictive of their behaviour in relevant field settings. For instance, people who tend to contribute more in public goods games are more likely to participate in local and national accountability institutions (Barr et al., 2014). Fishermen who exhibit more cooperation in a laboratory public goods game also show more cooperative behaviour in a real world common-pool resource problem by employing more sustainable fishing techniques; they use buckets with larger holes such that younger shrimps are not yet caught (Fehr & Leibbrandt, 2011). Another study (Rustagi et al., 2010) shows that Ethiopian communities that face serious common-pool resource problems are better able to maintain the commons if they have a higher share of people that display conditional cooperation in a public goods experiment. This study also provides evidence suggesting that causality runs from conditional cooperation to better maintenance of the commons resource. Behaviours consistent with conditional cooperation are also observed in field experiments (Keizer et al., 2008).

2 Can social norms explain cooperation-related behaviours?

All the abovementioned regularities are largely incompatible with the pure self-interest model, that is, they cannot be explained if it is common knowledge that all actors are rational and selfish. If free-riding is the dominant strategy at each contribution stage, there is also no incentive to enact costly punishment/rewards to induce cooperation, and reassortment or communication will not be effective either.

However, many of these regularities can, at least in principle, be explained if one directly assumes that a significant share of individuals has a desire to comply with a social cooperation norm (Ostrom, 2000). We call this the direct social norms approach (Bicchieri, 2006; Kimbrough & Vostroknutov, 2016; Lindbeck et al., 1999; Krupka & Weber, 2013) because it directly assumes (i) the existence of a norm c^* that is defined in terms of a specific behaviour and (ii) that individuals have an intrinsic desire to comply with c^* without providing a deeper micro-foundation of c^* and motives for norm compliance. In the context of cooperation, c^* describes the smallest cooperation level that is consistent with the normative prescription. Formally, this can be modelled by a utility function u_i in which individual i's utility depends positively on i's own material payoff x_i (which depends on all players' choices) while negative deviations of i's behaviour c_i from the

social norm c^* ($c_i < c^*$) generate some disutility:

$$u_i = \begin{cases} x_i - \gamma_i(c_i - c^*)^2 & \text{if } c_i < c^* \\ x_i & \text{if } c_i \geqslant c^* \end{cases}$$

The term $\gamma_i(c_i - c^*)^2$ denotes the psychic cost of deviating from the social norm (for simplicity these costs increase quadratically with negative deviations from the norm ($c_i - c^*$) and $\gamma_i \geqslant 0$ captures an individual's strength of the desire to conform to the norm. This approach represents a simple theory of conformism based on the assumption that negative deviations from the norm are, for some reason, psychologically costly for individuals with a strictly positive γ_i. In the context of cooperation, higher individual cooperation levels c_i are costly and thus reduce the individual's material payoff x_i but if c_i is below the norm c^* an increase in c_i reduces the costs of non-conformity $\gamma_i(c_i - c^*)^2$. For a sufficiently large level of γ_i the individual has therefore an incentive to obey the social norm c^*. Note that we assume for simplicity that positive deviations from the norm c^* have no psychological costs or benefits.

It is almost surely the case that the psychological cost of negative deviations from c^* (i.e., the γ_i's) vary across people but the assumption that there are some psychological costs of negative deviations makes sense in the light of the definition of a social norm because that definition implies that group members widely approve of the norm and that this is known by the subjects. Thus, subjects know that if they violate a social norm they are likely to face the disapproval of other people and for some people even the mere thought that others might disapprove of their action could constitute a psychological cost. In principle, γ_i could also represent the cost of deviating from a behavioural habit acquired in social life. Or the psychological cost of noncompliance could positively depend on how widely the norm is shared among the group members. However, in the following we assume for simplicity that γ_i is fixed and varies across individuals.

Unconditional normative prescriptions like "be selfless", "do the right thing" or "be moral" cannot explain the behavioural regularities described above. For example, they can neither explain communication effects (regularity 1) nor can they explain the decline in cooperation over time (regularity 2) or the higher levels of cooperation in partner compared to stranger matching (regularity 3). In contrast, a social norm of conditional cooperation can help explain all regularities but those described in regularities 6-8. This norm prescribes full cooperation as long as other group members also cooperate fully but if others' average cooperation becomes smaller it is normatively justified to match this reduction, that is, the conditional cooperation norm prescribes to contribute at

least as much as others' average contribution. Note that this implies that subjects' empirical beliefs about others' average cooperation become an important determinant of their cooperation levels—the more others cooperate the higher is the incentive to cooperate for an individual with a positive γ_i which explains regularity 2.

But this norm can also explain regularity 1: subjects with a very small γ_i ($\gamma_i \approx 0$) will defect while those with a sufficiently large γ_i and a high expectation about others' cooperation will cooperate in one-shot social dilemmas. Moreover, under face-to-face communication subjects often promise to each other to cooperate (Bicchieri, 2002) which is very likely to increase beliefs about others' cooperation. This increase in others' expected cooperation will then induce individuals with a sufficiently positive γ_i to increase their cooperation levels.

It has been shown (Fehr & Fischbacher, 2003; Fischbacher & Gächter, 2010) that the existence of imperfect conditional cooperators is the key ingredient for explaining regularity 3—the decay of cooperation over time in finitely repeated games. Conditional cooperation is imperfect if an individual does not match other group member's average cooperation perfectly but cooperates somewhat less than others are expected to cooperate on average. The above utility function assumes that people care positively for their own payoff and, therefore, individuals with a positive yet sufficiently low γ_i will not obey the norm c^* perfectly but reduce c_i somewhat below c^*, which implies imperfect conditional cooperation. However, if many individuals cooperate less than what each of them expect others to cooperate, jointly their expectations are too optimistic, which results in a downwards revision of their expectations and this then leads—via conditional cooperation—to a further decline in their cooperation rates, etc.

The existence of a conditional cooperation norm can also explain regularity 4—the higher cooperation rates under a stable group composition—and regularity 5, the existence of a framing effect on cooperation in the simultaneously played PD but not in the sequentially played PD (Ellingsen et al., 2012). When there is a stable group composition, even selfish individuals (i.e., those with $\gamma_i \approx 0$) have temporarily a strong incentive to cooperate because this generates benefits in future periods by inducing conditional cooperators to keep contributing (regularity 4). To explain regularity 5, recall that if there is a norm of conditional cooperation subjects who derive disutility from norm violations adjust their cooperation level to what they believe the other player will do in the simultaneously played PD. For optimistic beliefs they cooperate, for pessimistic beliefs, they defect. Under the plausible assumption that the label "Community Game" renders beliefs about the partner's cooperation more optimistic, conditionally cooperative subjects

will cooperate with higher frequency. However, for the second mover in the sequential PD beliefs are irrelevant because this player already knows exactly what the first-mover did. Thus, the frame can no longer change beliefs and therefore becomes irrelevant; and if a rational first mover anticipates the absence of a framing effect (s)he has no reason to condition behaviour on the frame either. Note that this explanation does not assume that the conditional cooperation norm changes across frames or between simultaneous and sequential play. The conditional cooperation norm can also explain why the addition of mutual reward opportunities to a public goods game increases cooperation (regularity 9). In the presence of mutual reward opportunities subjects can observe the cooperation level of other group members in the public good game after which they can spend money on rewarding other group members that costs them less than it benefits the rewarded subjects. This basically boils down to the opportunity of playing another bilateral prisoners' dilemma (PD) with each of the other group members after they observed others' cooperation levels. Obviously, the norm of conditional cooperation also applies to these PD games and because cooperation in the public good game can serve as a signal of cooperative intent, cooperation in the public good game fosters the belief that an individual will also cooperate in the PD. Therefore, mutual reward opportunities increase the incentive to cooperate in the public goods game.

Finally, the conditional cooperation norm can also help explain regularity 10, that is, why the assignment of cooperative individuals to the same group may cause high and stable cooperation. In terms of the direct social norms approach, cooperative individuals may be viewed as those with a sufficiently high γ_i such that for them perfect obedience with the norm ($c_i = c^*$) becomes optimal. If, in addition, these subjects are told that they are grouped together with other cooperators (Gächter & Thöni, 2005) they start with high expectations that trigger high cooperation which confirms the initial high expectation. The publicly known sorting of cooperative individuals into a group thus renders cooperation an equilibrium outcome.

For a similar reason, the existence of a conditional cooperation norm may also explain why cooperative advice by a previous generation of players that is made common knowledge among all current group members (regularity 10) causes large increases in cooperation rates. Cooperative advice that is common knowledge induces a general increase in the expected cooperation of other group members (Chaudhuri et al., 2006). Together with the norm of conditional cooperation the increased expectations then give rise to a general increase in cooperation rates.

However, there are of course other motives—such as equity or reciprocity motives—that

make similar predictions to those described above. Moreover, a norm of conditional cooperation cannot explain why subjects punish free-riders (regularity 6) nor subjects' preferences for playing the public goods game in an environment that allows for peer punishment (regularity 8). This follows simply from the fact that the conditional cooperation norm is defined in the space of cooperation behaviour and not in the space of punishment behaviour. One may, of course, stipulate the existence of another norm that renders punishment of free-riders a socially desirable act but (i) there is little evidence for this and (ii) it shows one of the drawbacks of an unconstrained direct social norms approach. By stipulating that a particular behaviour constitutes a social norm it is possible to explain any behaviour which renders such an approach irrefutable and thus empty—a problem that we take up later.

In real life, peer punishment ranges in severity and costliness from a simple raised eye brow to a hurtful smile, from outright ridicule to ostracism and the expulsion from social groups. Nevertheless, punishments in the lab capture key features of real life sanctions and also teach us that a significant share of participants will enact punishment systematically, even when it is costly and there is no personal material benefit for them.

3 The psychology of norm compliance

To make progress in understanding the potential impact of social norms on human cooperation, it is important to examine more closely the psychological reasons that induce individuals to comply with social norms. The direct social norms approach stipulates a normative behavioural standard and a psychological cost of non-compliance but does not provide a microfoundation for the behavioural standard and is typically not very explicit about the psychological cost of non-compliance. In principle, these costs could arise because individuals may be averse to actual, anticipated or merely imagined disapproval when deviating from the norm. In this case, compliance rests on an internalized desire for conformism, which has been challenged long ago as a general and sufficient basis for norm compliance (Wrong, 1961).

Another reason for psychological costs of norm compliance arises if individuals have an intrinsic desire for equity or fairness and social norms play a role in defining what is perceived as equitable or fair (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; López-Pérez, 2008). This case is also methodologically interesting because it implies that a collective phenomenon—the social norm—substantively affects the content of individuals' motivation by influencing what is perceived as fair, while the intrinsic desire for fairness

then ensures compliance with the norm. A third reason for costs of deviating from the social norm could be that individuals have a desire to reciprocate the behaviour of relevant others (Rabin, 1993; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006). In this case, the reciprocity motive applies, that is, the tendency to reward kind intentions with kindness ('positive reciprocity') and to punish hostile or unkind intentions ('negative reciprocity'). Note, however, that this motive requires a definition of what constitutes kind and unkind behaviour, which is typically also based on some normative notion of fairness/equity. For a reciprocally motivated individual, psychic costs of non-compliance arise, if the individual fails to reciprocate to a kind act with kindness or does not retaliate to a hostile act with a hostile response. Therefore, as in the case of fairness/equity motives, the reciprocity motive becomes operative on the basis of what is perceived as fair/kind and unfair/unkind.

A fourth reason for psychic costs of non-compliance arises if individuals have a propensity towards guilt aversion (Battigalli & Dufwenberg, 2007; Dufwenberg et al., 2011; Dhami et al., forthcoming). This theory rests on the idea that individuals experience the aversive, utility-decreasing emotion of guilt if they disappoint others. A social norm only exists if group members widely approve of the norm, and if there is widespread compliance then an individual act of non-compliance is almost surely disappointing other individuals. For example, if a subject believes that her partner in the prisoners' dilemma expects her to cooperate, then she disappoints him/her if she defects, and if the subject feels guilt and anticipates this emotion, she has an incentive to cooperate. Therefore, to the extent to which social norms generate the belief that others expect the individual to comply—a very likely belief in the presence of widespread compliance—a guilt-averse individual has some incentive to cooperate. However, if a social norm is systematically violated, such that the individual does not face a general expectation of compliance, a guilt-averse individual has no reason to comply with the norm. Guilt aversion is thus likely to generate conditional norm compliance behaviour that is mediated by individuals' beliefs about what others expect from them.

Finally, self-image theory assumes that individuals assign an intrinsic value to their self-image as a prosocial individual (Benabou & Tirole, 2011a). In this case, non-compliance with socially beneficial norms is detrimental for their self-image and provides a psychological deterrent for non-compliance. Similar to the case of fairness and reciprocity theories, this approach rests on some pre-existing notion—the notion of 'prosociality'—which is likely to be shaped by social norms.

It is interesting that all the abovementioned approaches rest on assumptions about in-

dividuals' intrinsic motivational properties. These motives—for example, the desire for fairness—are assumed to be stable across contexts. Stability in the desire for fairness does not mean, however, that the content of what is defined as fair is stable across contexts. It only means that individuals' preferences for implementing what is defined as fair, that is, their willingness to pay to implement the fair action, is stable while what is defined in a given society or group as fair or prosocial can be malleable. Thus, a main difference between social preference theories of equity, reciprocity, guilt aversion, and self-image and the direct social norms approach is that these theories are concrete about the motivational basis of norm compliance and the motives are assumed to be stable across contexts whereas the direct social norms approach remains vague with respect to the motives underlying norm compliance.

For example, both conditionally cooperative behaviour and the willingness to punish free-riders in a public goods game can arise from a desire for fairness or reciprocity. In other words, inequity-averse subjects and reciprocity-motivated subjects are often conditional cooperators as well as punishers (Fehr & Schmidt, 1999; Falk & Fischbacher, 2006) and, therefore, these motives contribute to the explanation of all the major qualitative regularities mentioned above (except the existence of antisocial or perverse punishment, which we discuss below). Likewise, the communication effects (regularity 1) as well as the framing effects (regularity 5) can be explained by stable preferences for equity or reciprocity because these preferences imply conditionally cooperative behaviour such that if frames and pre-play communication renders expectations about others' cooperation more optimistic, subjects will cooperate more.

Or take, for example, regularity 4 that 'partners' generally cooperate more than 'strangers'. The theory of inequity aversion or reciprocity can explain this finding by the regularity that the existence of inequity-averse or reciprocal subjects generates incentives for selfish individuals in a partner treatment to invest into cooperation during the early periods of a finitely repeated game (Ambrus & Pathak, 2011). This investment is profitable because it maintains the cooperation of the inequity-averse or reciprocal subjects in future periods. However, this incentive is absent in a stranger treatment where all interactions are one-shot so that there are no future gains. Note that this theory also explains that in a partner treatment, cooperation declines over time but restarts again if subjects play another finitely repeated game (Ambrus & Pathak, 2011). And because the theories explain why people punish free-riders, they can account for the punishment-related regularities 6–8.

In summary, social preferences for fairness/equity, reciprocity or a prosocial self-image

and the desire to avoid guilt are likely to play an important role in norm compliance. They provide an intrinsic motive to obey the normative standard to some extent and/or to sanction those who violate it. All of these theories are consistent with the notion that emotions are a key driver of the social preference although—with the exception of guilt-aversion theory, which models the emotion of guilt—they do not explicitly incorporate emotions in the model.

Although social preferences help in achieving norm compliance, it is important to distinguish them conceptually from social norms, which are defined as widely shared and approved normative standards. These standards are the essence of a social norm and they affect social preferences by defining what is considered as fair/equitable, kind or prosocial but they are conceptually nevertheless distinct. The direct norm approach is silent about the underlying motives that induce individuals to comply with a prevailing social norm and theoretical papers that apply this approach (Lindbeck et al., 1999) often make ad hoc assumptions about the social norm while empirical studies do not define ex ante the content of the normative standard but instead measure the norm empirically (Krupka & Weber, 2013; Krupka et al., 2016). This renders the direct norm approach more flexible and more difficult to refute unless it is possible to reliably identify the normative standard empirically over the relevant range of situations.

4 How can we identify social norms?

There are several methods for the identification of social norms (Fehr & Fischbacher, 2004b; Krupka & Weber, 2013; Cubitt et al., 2011; Reuben & Riedl, 2013; Bicchieri, 2017). One method builds on the premise that humans are willing to incur personal costs to sanction the violation of a norm even if they are not directly hurt by the violation. One reason for this willingness may be that norm violations have been shown to cause indignation or even outrage (Fehr & Gächter, 2002; Xiao & Houser, 2005; Bosman et al., 2005) aand these emotions may provide the raw material for the willingness to punish. Another reason may be that norm violators are typically perceived to deserve punishment (Carlsmith et al., 2002) and, therefore, sanctioning them provides satisfaction—a hypothesis that is consistent with the finding that reward-related brain areas are activated during the punishment of norm violators (DeQuervain et al., 2004) and that already preschool children and chimpanzees are willing to pay for watching the punishment of antisocial actors (Mendes et al., 2018).

Whatever the precise reason may be, if norm violations trigger the desire to punish

the perpetrators, we have a potential tool for identifying the norm as part of those behaviours that are not punished by uninvolved third parties. Various studies have therefore employed a third-party punishment paradigm for the study of social norms (Fehr & Fischbacher, 2004b; Henrich et al., 2006; Marlowe et al., 2008; Lewisch et al., 2011; Lergetporer et al., 2014). In these experiments, third parties more readily punish those who free-ride against a cooperative partner compared with bilateral defectors or cooperators, providing evidence for a norm of conditional cooperation (Fehr & Fischbacher, 2004b; Carpenter & Matthews, 2012; Kamei, 2017). Survey studies confirm that participants judge defection against a cooperative partner more harshly than mutual defection (Cubitt et al., 2011).

An important method for the identification of social norms is based on the idea that social norms provide a focal point such that subjects' normative judgements are coordinated on this focal point (Krupka & Weber, 2013). This approach provides an incentivized measure of social norms by asking subjects to rate the extent to which an action is 'socially appropriate and consistent with moral or proper social behaviour'. Subjects are not asked to provide their own personal evaluation, but to indicate what they believe is the most common answer, and they earn a monetary reward if their rating coincides with the modal answer of others. This method has already been employed in several studies to elicit the social norm in social dilemmas (Kimbrough & Vostroknutov, 2016; Gächter et al., 2013) and in one of these studies (Kimbrough & Vostroknutov, 2016) identifies a conditional cooperation norm in the public good game.

One study (Bartling & Özdemir, 2017) applied this method to measure whether the punishment of unfair proposers in the ultimatum game—by rejecting their offer—is a social norm. Interestingly, the study shows that this is clearly not the case. We conjecture that this is also likely to hold in social dilemma situations, suggesting that the desire to punish free-riders derives from other motives such as to avoid inequity (Fehr & Schmidt, 1999; Dawes et al., 2007) or to reciprocate to unfair actions (Falk & Fischbacher, 2006; Carpenter & Matthews, 2012).

Another method for the identification of social norms in social dilemma games has recently been presented in two papers (Fehr & Schurtenberger, 2018a; Fehr & Williams, 2018). Here, each subject of the group is asked to indicate what other group members should contribute to the public goods. The average of subjects' normative requests is afterwards conveyed to all group members and is likely to constitute a general normative standard of cooperation because it is commonly known and reflects the group members' views. Moreover, the higher subjects' agreement in their normative requests, the more

the average request will constitute a legitimate normative standard (Fehr & Schurtenberger, 2018a). One advantage of this method is that it can be easily implemented in every period of a public goods game such that the level and the strength of the norm can be identified continuously. Also, this method supports the existence of a conditional cooperation, that is, the average requested contribution in a period is declining in subjects' average actual contributions in the previous period. In addition, the data show that when direct targeted punishment of free-riders is possible, subjects strongly obey the average normative request in their actual cooperation choices (Fehr & Schurtenberger, 2018a).

hus, taken together, there is ample and diverse evidence for the existence of a conditional cooperation norm in social dilemma situations while there is little or no evidence that punishment of free-riders constitutes a social norm. These results show that one can provide discipline to the direct social norms approach and they strengthen the conjecture that a conditional cooperation norm shapes human cooperation. However, these norm elicitation approaches do not yet prove that cooperation behaviour is causally affected by social norms because they—so far—only establish a correlation between the social norm and actual cooperative behaviour (Kimbrough & Vostroknutov, 2016).

5 Do social norms causally affect cooperation behaviour?

The potential causal effect of social norms on behaviour has been studied in various ways. A prominent approach (Cialdini et al., 1991; Kallgren et al., 2000) assumes that social norms need to be activated, that is, become the focus of subjects' attention to affect behaviour. Based on this view, a causal effect of social norms can be identified by varying the salience of the norm with various priming techniques. This literature shows that when subjects' attention is shifted towards social norms they begin to act in a more norm-congruent way (Cialdini et al., 1991; Kallgren et al., 2000; Berkowitz & Daniels, 1964; Berkowitz, 1972; Hallsworth et al., 2017). For example, in one study (Kallgren et al., 2000), , car drivers, who did not know that they were part of an experiment, saw the following handbill on their windshield: "April is Keep Arizona Beautiful Month. Please Do Not Litter". In a second condition, the text on the handbill was "April is Conserve Arizona's Energy Month. Please Turn Off Unnecessary Lights" and in a third (control) condition they could read "April is Arizona's Fine Arts Month. Please Visit Your Local Art Museum". In line with the hypothesis that a stronger activation of the anti-littering norm leads to less littering, car drivers threw the handbill on the ground in only in 10% of the cases in the first treatment, and in 18% and 25% of the cases in the second and third conditions, respectively. Findings like these raise the question of

which aspect of the social norm is the causal driver of the behaviour change. Does the increase in the salience of the norm change the social appropriateness rating of norm-compliant behaviour? Or does it merely change subjects' views about how widely the norm is shared? Or does it change subjects' feelings of guilt if they litter? Unfortunately, we do not know the answer to these questions.

The above-mentioned method for norm identification (Fehr & Williams, 2018; Fehr & Schurtenberger, 2018a) through individual normative requests can also be used to study the causal impact of social norms on behaviour. In treatments with normative requests, the average request constitutes a commonly known standard of behaviour that is absent in treatments without normative requests. In one study (Fehr & Schurtenberger, 2018a), the authors introduce the norm formation opportunity in finitely repeated public goods games where the possibility to punish other group members is either absent or present. Interestingly, when the possibility of punishment is absent, the opportunity to form a normative standard has no impact on behaviour while in the presence of the possibility to punish, the normative standard causes a significant and stable increase in cooperation rates (Fig. 3).

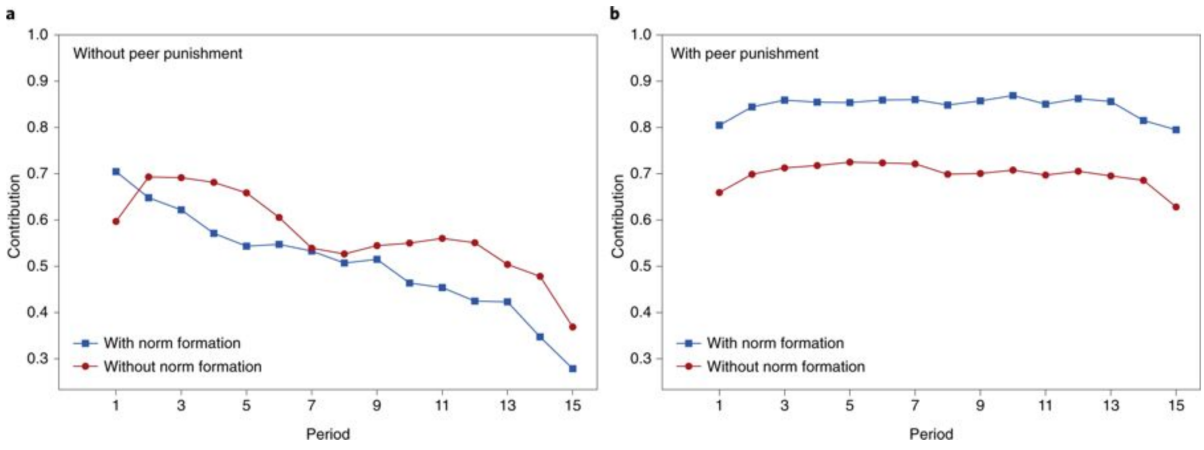


Figure 3: The effect of social norms with and without punishment (Kallgren et al., 2000)

Fig. 3. Average normalized contributions over time (1 = full contributions; 0 = complete free-riding) in fixed groups of four subjects that play a public goods game for 15 periods. **a**, Treatments without punishment. **b**, Treatments with punishment. Treatments with a punishment opportunity allow for the counter-punishment of those who punish free-riders to examine whether norms have a causal impact in an environment that has been shown to be hostile for human cooperation (Nikiforakis, 2008).

This radically different impact of social norms on cooperation when there are punishment opportunities exists despite the fact that the normative standard in the punishment and no-punishment treatment is very high and statistically indistinguishable during the first three periods. Nevertheless, substantial norm deviations occur in the absence of punish-

ment from the very beginning while in the presence of punishment the norm is largely obeyed throughout the whole experiment. Thus, the existence of a normative standard that renders high cooperation the socially most appropriate action, and focuses attention on the normative standard, is per se not sufficient to induce a change in cooperation behaviour, suggesting that intrinsic motives for norm compliance are not sufficiently strong and that the punishment threat is needed to establish a stable norm-driven behaviour change in a population of heterogeneously motivated actors.

6 Normative constraints and peer punishment (in)efficiency

The existence of punishment opportunities in public goods games causes strong cooperation increases in many, but not in all, cultures (Herrmann et al., 2008; Gächter & Herrmann, 2009, 2011). In particular, in those countries that have weak norms of civic cooperation—defined as the willingness to evade taxes, make fraudulent claims to receive welfare state benefits or dodging fares on public transport—the antisocial punishment of cooperators is particularly strong and is associated with detrimental effects on overall cooperation rates. This finding is consistent with the view that norms of civic cooperation have a causal, constraining effect on antisocial punishment. However, the finding does not prove causality because there could be other reasons that may account for the correlation between antisocial punishment and norms of civic cooperation. For example, countries with low norms of civic cooperation often also have bad schools (for example, because of teacher absenteeism or low teacher quality (Hanushek & Woessmann, 2016; Hanushek & Rivkin, 2012)) and school or teacher quality might shape both norms of civic cooperation and restraints on antisocial punishment.

Although the antisocial punishment of above-average cooperators by those who cooperate less tends to be rare in Western cultures, it has been observed from the beginning and several potential reasons for its existence have been mentioned (Fehr & Gächter, 2000a). First, in rare cases, it may simply reflect a random choice error. Second, there is evidence that a small, yet significant proportion of subjects regularly displays envious or spiteful motives (Fehr et al., 2008; Bruhin et al., forthcoming), implying that they prefer to spend money to hurt others regardless of their level of prosociality. Third, antisocial punishment may be the result of a coordination failure among reciprocally motivated subjects that are in principle willing to cooperate. Consider a reciprocal subject with pessimistic beliefs about others' cooperation. These subjects may cautiously start with an intermediate or low level of cooperation while other subjects have optimistic expectations, start with high cooperation and punish those who cooperate less. The pessimistic, yet willing, low

contributor may view this as an unfair punishment and may thus retaliate in the next period against the high contributors. These events may spoil the whole group and lead to a process of punishment and counter-punishment with detrimental effects on cooperation. In fact, if subjects are given explicit counter-punishment opportunities (Cinyabuguma et al., 2006; Nikiforakis, 2008), some subjects use them to the detriment of the group's cooperation and welfare by punishing those who punished them for free-riding. More generally, public goods experiments that allow for peer punishment often fail to increase the overall welfare of the group members for an extended period of time despite the large increase in cooperation rates (Fehr & Gächter, 2000a; Gächter et al., 2008; Rand et al., 2009). The reason for this is the high collateral cost associated with peer punishment.

However, the very fact that peer punishment can get out of control suggests that societies have developed mechanisms to constrain and control it. After all, peer punishment is physically always possible when two or more individuals directly interact with each other. It appears impossible for society to ever control or constrain all the different forms of peer punishment—that range from a raised eye brow or verbal insult to mobbing, ostracism, public shaming and corporal punishment—except through the normative control of people's behaviour. The literature on simple societies (Wiessner, 2005; Mathew & Boyd, 2011) provides ample evidence of the ways in which societies impose constraints on punishment. One study (Wiessner, 2005), for example, reports how the Ju/'hoansi bushmen, a group of hunter-gatherers living in Botswana, exert peer punishment according to strong habitual and normative constraints. For instance, if a man is publicly criticized for norm violations, this is often done by a women to avoid the escalation of arguments among men.

Rather than rely on peer-to-peer sanctioning, individuals will often prefer some type of institutional arrangement to regulate punishment by either ruling out peer punishment completely (Sutter et al., 2010) or replacing it with a centralized state that automatically imposes taxes to finance public goods (Markussen et al., 2014) or by an enforcement mechanism that rules out antisocial peer punishment (Yamagishi, 1986; Ertan et al., 2009; Traulsen et al., 2012; Andreoni & Gee, 2012). But how is it possible to achieve this without also ruling out peer punishment altogether and more fundamentally, how is it ever possible to rule out peer punishment altogether in a world in which people socially interact with each other and in which the centralized legal enforcement of rules is always imperfect?

This question can be answered by comparing the punishment patterns in settings with and without the opportunity for normative requests (Fehr & Schurtenberger, 2018a). It turns

out that when subjects can form a normative cooperation standard, the punishment of free-riders becomes less severe. Thus, the normative standard increases cooperation while simultaneously decreasing the punishment of free-riders, suggesting that the punishment of free-riders becomes more effective. In fact, punished free-riders indeed increase their cooperation subsequently more strongly when the normative standard is present (Fehr & Schurtenberger, 2018a). Antisocial punishment also decreases in the presence of a normative cooperation standard, thus lending support to the hypothesis that norms of civic cooperation may causally reduce antisocial punishment.

Despite the high potential collateral cost of normatively unconstrained peer punishment, it has been observed that participants will prefer this over a setting with no opportunities for targeted punishment (regularity 8). However, if subjects additionally can migrate to normatively coordinated peer punishment and normative coordination and punishment by a central authority, participants never enter the uncoordinated peer punishment setting. The institutions with normative coordination minimize or fully eradicate antisocial punishment and generate high levels of cooperation without the collateral damages associated with uncoordinated peer punishment (Fehr & Williams, 2018). This demonstrates that the traditional uncoordinated peer punishment institution fails to capture a very important dimension: the strong demand for normative coordination and regulation—a demand that societies who inevitably have to rely on some forms of peer sanctioning typically satisfy through the formation of social norms that put constraints on individuals' sanctioning behaviour. Of course, groups will not automatically solve inefficient peer sanctioning through informal constraints, but it seems likely that those groups who do solve this problem in a more efficient way will be more successful because they are better able to solve their collective action problems (Boyd & Richerson, 1994, 1992; Henrich, 2004). Therefore, they are better able to compete with other groups. Thus, conclusions regarding the effectiveness and the welfare properties of peer punishment may provide a misleading picture if they are based on institutional settings that rule out suitable normative consensus building opportunities that can put constraints on peer sanctioning.

7 Summary and open questions

The pervasiveness of social norms and the ubiquity of cooperation among non-kin are two salient features of human societies. Many social norms are beneficial for overall society and compliance with them can be viewed as acts of cooperation. Although humans are by no means the only species displaying cooperation among individuals, it has often been pointed out that the breadth and depth of human large-scale cooperation among non-kin

in a globalized world, as well as the observed cooperation in one-shot encounters, appear unique in the animal kingdom (Fehr & Fischbacher, 2003; Hammerstein, 2003; Stevens & Hauser, 2004; Boyd & Richerson, 2005). Several potential factors—such as limited memory or excessive time discounting (Stevens & Hauser, 2004; Stephens et al., 2002)—may constitute evolutionary obstacles to cooperation in animal species but perhaps the cognitive prerequisites for social norms are also relevant. For example, the very notion of a normative standard—what ought to be done—is rather complex and perhaps even impossible to identify reliably in species that lack sophisticated language. The same applies to the notion of normative approval and disapproval. Therefore, it is perhaps not surprising that our closest living relatives do not seem to share some of our most fundamental norms of fairness and cooperation (Jensen et al., 2007a,b; Ulber et al., 2017) (although see Proctor et al. (2013); Brosnan et al. (2005)) and that there seems to be no evidence for third party punishment of norm violations harming non-kin in non-human species (Riedl et al., 2012). In contrast, third party punishment of non-kin and even strangers is frequent in humans (Fehr & Fischbacher, 2004b; Henrich et al., 2006; Jordan et al., 2016) and young children already have a working knowledge of social norms (Ulber et al., 2017; McAuliffe et al., 2015; Cummins, 1996). The widespread prevalence of social norms may therefore well be one of the defining characteristics of our species and a crucial determinant of human cooperation.

The evidence suggests that human cooperation is strongly affected by normative considerations. Various methods indicate the existence of a strong conditional cooperation norm. The behavioural strength of the conditional cooperation norm probably also derives from its relation to principles of equity and reciprocity. Compliance with social norms relies on the existence of social preferences that incorporate abstract normative principles such as equity or reciprocity—which also provide foundations for the willingness to punish norm violators—or are based on the desire for avoiding disapproval, a prosocial self-image or the avoidance of disappointing others. Social norms also appear to guide and constrain punishment behaviour and subjects have a strong desire for environments that enable normative coordination.

There are, however, still many important unanswered questions. Reliable empirical knowledge about the precise channels through which norms have a causal impact is, for example, still scarce. Does the normative standard shape behaviour directly via an intrinsic utility component or does it have an impact by affecting and coordinating beliefs about others' cooperation. Or does it guide the punishment of free-riders and affect beliefs about punishment in case of non-compliance? In addition, there are many other intriguing and exciting questions that are awaiting an answer (see important unresolved

research problems), implying that there is still much to discover in this area of research.

Important unsolved research problems

1. What are micro-sociological and psychological processes that facilitate and hinder the development of a social norm?

2. What is—at the conceptual level—the precise relationship between social preferences and social norms and how can we distinguish them empirically? How do social norms influence the motivational content of social preferences and, for given social preferences, how do they affect compliance with normative standards?

3. What determines individuals' agreement with the "ought component" of norms (Fehr & Schurtenberger, 2018a)? How do they come to internalize or reject a normative standard?

4. What explains the formation and the decay of social norms and how can we explain changes in the normative content, i.e., the "ought component" of social norms (Fehr & Schurtenberger, 2018a)?

5. What are the long-run environmental and economic determinants of social norms (Henrich et al., 2010; Alesina et al., 2013; Ellickson, 2001; Lowes et al., 2017)? And how do normative standards evolve in the context of conflicting economic interests (Reuben & Riedl, 2013)?

6. How do economic incentives, the human desire for social approval and normative standards interact? When are they complements and when do economic incentives undermine normative standards and approval incentives (Benabou & Tirole, 2011b)?

7. How does actual compliance and non-compliance shape the development of normative standards (Fehr & Schurtenberger, 2018a)?

8. Through which interventions and public policies is it possible to shape social norms (Bicchieri, 2017) and which aspect of the norm and norm-related behaviours—the content of the normative standard, social agreement with the normative standard, behavioural compliance with the standard—is changed by the intervention?

9. How do legal institutions—apart from their sanctioning capacity—affect social norms and how do social norms affect the effectiveness of legal institutions (Benabou & Tirole, 2011b; Posner, 2000)? To what extent do legal institutions shape

normative standards by setting precedent, fall back rules or through expressing what is normatively approved and expected (Sunstein, 1996)?

10. To what extent and in which ways do social norms influence important economic and social patterns (Fehr & Williams, 2018; Fehr & Schurtenberger, 2018a; Akerlof, 2007; Allcott, 2011; Nolan et al., 2008)?

Chapter III

The Superiority of Decentralization in Social Norm Enforcement

The Superiority of Decentralization in Social Norm Enforcement

Ernst Fehr, Yagiz Özdemir & Ivo Schurtenberger

Abstract

The resolution of social dilemmas is a fundamental problem of societies. One of its remedies is the enforcement of a social cooperation norm. Monitoring of actions is one of its prerequisites, however, the available signals are generally imperfect, and therefore leave room for errors in enforcement. Two types of errors can occur, which both are detrimental to cooperation. Norm followers may wrongfully receive some punishment (Type-I) and norm violators may elude it (Type-II). We design a laboratory public goods game where signals about actions are imperfect and either exogenous public, exogenous private or endogenous private. In these three information environments we study human behavior and the relative performance of two different enforcement institutions; decentralized peer-to-peer punishment is compared to a centralized regime. Decentralization achieves significantly higher cooperation rates than centralization under all three monitoring conditions. There is a trade-off between Type-I and Type-II errors of punishment, where decentralization fares worse regarding the former and much better regarding the latter. Moreover, we find substantial demand for additional signals about the contribution decisions of other group members. This behavior substantially reduces the prevalence of Type-I errors, which boosts cooperation rates in turn. Finally, we show that private signals, compared to public signals, are a curse for peer-to-peer punishment, despite the fact that overall more information is available in this setting.

JEL classification: C92; H41; D23

Keywords: Public goods; Norm enforcement; Imperfect monitoring; Private monitoring; Institution; Decentralized; Centralized; Public monitoring; Information acquisition; Punishment errors

Citation: Fehr, E., Özdemir, Y., & Schurtenberger, I. (2018). The superiority of decentralization in social norm enforcement. *Working Paper.*

---

1 Introduction

Social dilemmas constitute one of the fundamental problems that societies need to solve to achieve efficient outcomes. According to Ostrom (1998) "social dilemmas are found in all aspects of life, leading to momentous decisions affecting war and peace as well as the mundane relationships of keeping promises in everyday life" (p. 1). Social dilemmas arise whenever a Pareto optimal outcome can be achieved if group members cooperate, but individual payoff-maximizing choices lead to Pareto inferior allocations. A canonical example for a social dilemma is the public goods game, where the social optimum is achieved if all agents contribute to the public good, but each agent has an incentive to free ride on the contributions of other group members.

Empirical research has shown that cooperation in public goods games can be sustained if (costly) punishment options exist, since subjects are willing to bear costs to punish group members who violate the social norm of cooperation (Fehr & Schurtenberger, 2018b), whereas cooperation typically breaks down if sanctioning mechanisms are not available (e.g. Yamagishi, 1986; Ostrom et al., 1992; Fehr & Gächter, 2000a).

The literature on the enforcement of cooperation in public goods games has largely been based on a simplifying assumption about the underlying monitoring technology, that is, the assumption that all actions of other group members are perfectly observable before punishment decisions are made. Recently, this assumption has been relaxed to allow for imperfect signals about the contribution decisions of other group members, leading to monitoring environments where all agents receive noisy signals (e.g. Ambrus & Greiner, 2012, 2015; Grechenig et al., 2010; Fischer et al., 2013; Nicklisch et al., 2015).

We argue that a realistic monitoring structure produces signals with the following three properties. First, signals are *imperfect*, that is, actions are not perfectly observable, and therefore cooperators might be mistaken for defectors and vice versa. Second, signals are, at least to some degree, *private* in nature. By this we mean that signals about someone's actions need not to be the same for everyone, certain members of a group might perceive someone as a cooperator, whereas others' signals might suggest that this someone is in fact a norm violator. Third, information is *endogenous*. There often exist possibilities to seek further information about someone's true behavior by exerting some kind of effort.

Under imperfect monitoring there are two types of norm enforcement errors that can occur. A norm follower might receive some form of sanction for a violation she did not commit. This constitutes a "false positive" or Type-I error. On the other hand, norm violators might not be perceived as such, or the uncertainty about their true actions

potentially discourages the enforcement of cooperation norms, and they therefore elude punishment. This acquittal of norm violators constitutes a "false negative" or Type-II error. Both errors are detrimental to cooperation, first, by discouraging cooperators to further contribute to the common cause, and second, by missing the opportunity to enforce future cooperation of free-riders. Errors in the enforcement of the cooperation norm are determined by a complex combination of several interwoven factors including the information structure, punishment behavior, signal acquisition, and the enforcement institution.

In his seminal Leviathan Hobbes (2005 Orig. pub. 1651) argues in favor of a centralized authority, and questions the efficacy of self-governance to overcome what he regards as the state of nature: *bellum omnium contra omnes*.[1] Ostrom et al. (1992) reply "self-governance is possible" when a group of humans face a social dilemma situation. In this paper, we contribute to this perpetual discourse by illustrating the causal effects of the monitoring possibilities and the imminent errors in sanctions on human behavior and the relative performance of social norm enforcement institutions.

Let us examine the *potential* for enforcement errors under centralization and decentralization due to wrongful signals by moving from the simplified case of perfect information to the realistic case of imperfect endogenous and private signals. For the examination one needs to consider the information that is available to those who hold sanctioning power. Under centralization few or even a single actor comprise this entity. Under decentralization the power to sanction is dispersed among multiple actors. Clearly, when information is perfect, there is no potential for unintended errors in either institution. When there are public imperfect signals, the potential for errors of punishment is the same in both institutions because information is equal. But with independent (private) signals, the information base changes between the two institutions due to the varying number of actors with the possibility to sanction. Under decentralization, there is a greater chance of at least one correct signal, but also a greater chance of at least one false signal about a certain agent. This means, for cooperators and defectors alike, an increase in the potential for punishment, and therefore a greater potential for Type-I errors and a lower potential for Type-II errors. When imperfect and private signals can be improved, then the potential for errors is endogenous and depends on the information acquisition behavior of those who wield power over sanctions.

A profound understanding of human behavior under these settings is important, because designing effective mechanisms to overcome social dilemmas does not only involve picking

[1]lat. for *war of each against all*, war may be understood as competition or struggle in this context.

the punishment institution, but also shaping the monitoring capabilities. For example, organizations can increase work-flow transparency, making it easier for employees to monitor the behavior of peers or increasing the correlation between signals through more open communication. Furthermore, organizations can prioritize vertical control, where superiors monitor the behavior of employees, or horizontal control, where compliance with productivity targets and behavioral norms is monitored by peers (McAllister, 1995).

Such an examination is challenging, especially in a field setting, for the following reasons. First, the details of the monitoring structure are usually unobservable, for instance, a researcher does not observe the probability with which signals are correct or how strong the correlation is between individual signals. Second, it is unfortunately often impossible to verify someone's true actions. Hence, an individual cannot be classified as a cooperator or defector, which finally means that the rates of "false positives" and "false negatives" are unknown. Third, variation in monitoring and sanctioning institution are endogenous, which renders the establishment of causal relationships hard. For example, one can expect the adoption of a centralized punishment institution if such an institution has access to more accurate signals.

We circumvent these problems by using a controlled laboratory experiment featuring a public goods game with punishment at its core. This setting leaves the design of the monitoring and the punishment institution to the researcher, who can randomly assign subjects to treatment conditions, which in turn allows establishing causal effects of information structure and institution on key outcome variables of cooperation. In our experiment, subjects face a binary choice whether or not to contribute to the public good. This allows a clear classification of subjects as either cooperators or defectors resulting in measurable rates of enforcement errors.

Using a public goods game, we compare a decentralized peer-to-peer punishment institution to a centralized punishment regime, where all sanctioning power is concentrated in the hands of a single, randomly selected authority. We begin our examination with the most realistic case of endogenous private and imperfect signals. All group members, including the authority, receive a private signal about the binary contribution decisions of each group member, a signal that is with 90% probability correct and with 10% probability wrong. In this endogenous monitoring condition, subjects can costly acquire additional pieces of information. The results show that decentralization of punishment induces higher cooperation rates than centralization. There exists a trade-off between the two institutions regarding the two types of enforcement errors. The decentralized institution punishes many more cooperators, but has a distinct advantage in not making the

decisive mistake of letting defectors to elude punishment. The data reveals that subjects almost exclusively acquire additional signals when the initial signal suggests that a group member did not contribute to the public good. This information acquisition behavior suggests that subjects focus on avoiding the punishment of the innocent (Type-I error).

We conduct additional treatments to further investigate to what end subjects make use of the possibility to improve information. In these treatments we remove the possibility to acquire new signals and subjects have to make their decision based on their initial single signal. Subjects indeed manage to improve information to reduce the prevalence of "false positives." This is the case for either institution. Not only the error rate benefits from gathered information, but also the cooperation rate itself. Under both institutions, subjects contribute significantly more when information is endogenous compared to exogenous. Decentralization of social norm enforcement fares also better than centralization under this second information structure. One might argue that the advantage of decentralization only originates from the fact that aggregate information of those who hold punishment power is greater under this institution. This is the case because the authority receives a total of four signals about the peers, but the peers receive a total of twelve (each peer receives three) signals.

We conduct two more treatments featuring public signals instead of private ones: This way the information between the two institutions is held constant. The results from these treatments show that having private signals is not a blessing but rather a curse for peer-to-peer punishment. Cooperation rates under public signals are even greater than under private signals with peer punishment. Centralization does not profit from making signals public instead of private. The trade-off between the two error rates under private signals already indicates that this alleged advantage in information does not come without costs. More information also means a greater potential for some false information. Note that in our experiment public signals are not associated with greater credibility than private signals.

Taken together we conclude that decentralization is superior in sustaining high cooperation rates in all three imperfect monitoring conditions considered in this study, that is, under exogenous public, exogenous private and endogenous private signals. The data suggests that the lower rate of unpunished defectors is key for the dominance of peer-to-peer punishment, even though it comes at the cost of more punished cooperators.

This study is related to several strands of the economic literature. It was shown that both self-governed peer-to-peer punishment (e.g. Fehr & Gächter, 2000a) and a centralized authority (e.g. Baldassarri & Grossman, 2011) have the means to prevent a collapse

of cooperative behavior. However, several studies suggest that centralization of punishment might mitigate potential disadvantages of peer-to-peer punishment regimes. First, decentralized punishment institutions can be prone to antisocial punishment, that is, the intentional punishment of cooperators, rendering it difficult to sustain cooperation (Cinyabuguma et al., 2006; Herrmann et al., 2008; Gächter & Herrmann, 2009, 2011). Second, since agents in the decentralized system have incentives to free-ride on the altruistic punishment decisions of other agents (Fehr & Gächter, 2002), centralization might mitigate the second order public good problem of punishment. Third, coordinating the appropriate severity of sanctions might prove more difficult in a decentralized than in a centralized setting.

However, empirical studies comparing decentralized and centralized punishment institutions do not find evidence for a superior performance of centralization in terms of cooperation rates. The results of these studies are either based on perfect (O'Gorman et al., 2009) or exogenous public (Fischer et al., 2013) information. In case of Fischer et al. (2013), the information structure is different than the one considered in this study. Instead of a binary contribution decision, subjects can contribute an integer amount between zero and twenty in their design. A wrong signal then just depicts any other possible contribution level with equal probability. We show that in several more realistic monitoring environment cooperation rates are higher when sanctioning power is decentralized. Our findings imply that in order to achieve similar cooperation rates as in a peer-to-peer punishment setting, centralized punishment institutions need to be associated with additional features that are potentially beneficial for cooperation, such as a commitment to sanctioning rules (Putterman et al., 2011; Tyran & Feld, 2006; Andreoni & Gee, 2012) or the election of the authority by group members (Baldassarri & Grossman, 2011).

Also related to our study is the one by Nicklisch et al. (2015), which examines the emergence of centralized institutions and the role of imperfect information. Subjects vote for their institution (no punishment, decentralized or centralized) by feet. Centralized institutions only emerge under two conditions, first, the randomly selected authority has to refrain from punishing cooperators and second, there needs to be some degree of noise.

This paper also speaks to the strand of literature concerned with imperfect monitoring in social dilemma situations. Generally speaking, imperfect monitoring poses a challenge to achieve efficient outcomes in such settings. Carpenter (2007) varies the size and the number of observable members of groups in a public goods game. Imperfect monitoring, that is, in their case, when not all group members' contributions are known, leads to fewer contributions. In an infinitely repeated prisoner's dilemma, subjects' welfare was

found to decrease with the level of noise in public signals (Aoyagi & Fréchette, 2009). Grechenig et al. (2010) reveal a high willingness of subjects to punish in a public goods game with noisy signals. Furthermore, punishment in this setting cannot maintain high contribution, and even causes welfare to be lower than in a setting without punishment. Ambrus & Greiner (2012) show that with imperfect signals an increase in the severity of punishment does not monotonically increase contribution and welfare, which it does under perfect information. In Ambrus & Greiner (2015) a democratic punishment institution—subjects vote on the punishment of others—outperforms a decentralized peer-to-peer setting both under perfect and under imperfect public signals.

The observed focus of information acquisition behavior on reducing Type-I punishment errors is consistent with the findings of Dickson et al. (2009), who study behavior in public goods games with centralized punishment in different monitoring conditions. In the False Positives treatment, signals can only be inaccurate if the underlying true decision is to cooperate, whereas in the False Negatives treatment, signals can only be inaccurate if the true underlying decision is to defect. Dickson et al. (2009) find that, unlike in the False Negatives treatment, in the False Positives Treatment authorities are reluctant to use punishment, because they want to avoid the risk of punishing a cooperator. The reluctance to punish in the False Positives treatment is highly detrimental for cooperation, since defectors remain unpunished, leading to a so called "False Positives Trap". However, with a signal accuracy of only 60%, the monitoring technology used in Dickson et al. (2009) delivers only very limited information to the authority, causing high error probabilities of punishment. In our study, where the signal accuracy is 90%, this "False Positives Trap" seems to be less of a problem at least for peer punishment; many peers are willing to take the risk of punishing a cooperator in order enforce cooperation. However, subjects also make use of opportunities to reduce the risk of Type-I errors by acquiring further information before punishing an alleged defector.

Markussen et al. (2016) use a public goods game with exogenous (automatic) punishment, and therefore exogenous errors of punishment. Subjects in their study have a greater willingness to pay to prevent a Type-I error than Type-II error of the same magnitude, which is in line with our subjects' focus on avoiding Type-I errors. There is further evidence from a study by Feess et al. (2014) that uses a stealing game. Subjects who need to judge an alleged thief of a charity donation care more about Type-I than about Type-II errors.

The findings of this study are potentially interesting for the design of effective social norm enforcement institutions. Decentralization outperforms centralization per se under

realistic, less ideal, monitoring structures. Unless centralization is equipped with other performance enhancing features, a peer-to-peer institution seems to enforce cooperation with greater efficacy. Regarding the design of the monitoring capabilities decision makers should have the possibility to improve their information base about actions of other group members, that is, the costs of monitoring should be kept low. Decision makers are willing to incur costs to improve information in order to avoid Type-I errors which in turn fosters cooperative behavior. Finally, when some form of peer-to-peer punishment institution is in place, then signals should be as public as possible, because the private nature of signals is rather a curse for such a setting. Centralization on the other hand is indifferent between public and private signals.

The remainder of the paper is organized as follows. Next, we describe the general experimental design, then we outline our analysis and our results as well as describe our additional treatments, and finally, we conclude.

2 Experimental Design

Treatments are based on a linear public goods game with punishment, which is repeated for 25 periods. At the beginning of the experiment subjects are randomly allocated to groups of five, each group consisting of four Peers (P_1-P_4) and one Authority (A). Groups and roles remain fixed throughout the experiment, but the identification number of Ps is randomly assigned in each period to avoid reputation effects.

Every period has three stages: contribution (stage 1), monitoring (stage 2), and punishment (stage 3). While the contribution stage remains the same across all treatments, the other two stages differ depending on the treatment. We employ two different punishment and monitoring institutions, which defines who holds the power to acquire new signals and to exert punishment. In the decentralized institution (treatment END-PRI-DEC), this ability resides with the peers. In the centralized institution (treatment END-PRI-CEN), it is focused in the hands of a randomly selected authority. During the contribution stage Ps make a binary choice whether or not to contribute their endowment to the public good. Contributions are doubled and then redistributed to all group members. Each group member, including player A, receives one initial imperfect private signal about the contribution decisions of other Ps. Specifically, with 90% probability the signal corresponds to the actual contribution of the respective P and with 10% probability the signal provides wrong information. We introduce the possibility to costly acquire further signals, that is, Ps or A (depending on the institution) can spend money to reveal up to two

additional signals for each peer. Finally, Ps or A (again depending on the institution) have the possibility to costly exert punishment, that is, to spend 1 Token in order to reduce the target's income by 4 Token. To hold endowments and total costs constant across treatments, A and Ps share the costs of punishment and monitoring.

The following section describes the three stages and the differences between the treatments in more detail. Table 1 lists the implemented values of all parameters (1 Token = CHF 0.05).

Stage I: Contribution

In the contribution stage, which is the same in both treatments, each peer P_i receives an endowment of e^{PG} and decides whether to contribute the whole endowment to the public good ($c_i = 1$) or not ($c_i = 0$). Using a binary choice has several advantages for the purpose of this paper. First, is allows a clear distinction between *cooperators* and *defectors*, and therefore allows to identify punishment errors. A "false positive" (Type-I error) punishment errors occurs if a cooperator is punished. If a defector eludes punishment a "false negative" (Type-II error) punishment error occurs. Second, incorrect signals, due to imperfect monitoring, is meaningful and easily understandable. Defectors are sometimes mistaken as cooperators, and cooperators sometimes appear to be defectors. Since player A is passive in this stage and cannot contribute to the public good, they do not receive any endowment. Contributions to the public good are multiplied by M and distributed back equally to all five group members; hence, marginal per capita return of a contribution is given by $M/5$. Player A is included as a beneficiary of the public good, because otherwise the incentives for exerting punishment would not be comparable between the decentralized and the centralized punishment treatment. Parameter M is chosen such that the game constitutes a social dilemma—it is strictly dominant not to contribute to the public good, but if all peers defect, the resulting outcome is Pareto inferior compared to the case where all peers cooperate. The monetary payoff of P_i from stage I is given by

$$\pi^I_{P_i} = e^{PG}(1 - c_i) + \frac{M}{5}\sum_{j=1}^{4} e^{PG} c_j.$$

A's payoff is given by

$$\pi^I_A = \frac{M}{5}\sum_{j=1}^{4} e^{PG} c_j.$$

Stage II: Monitoring

In the monitoring stage subjects receive information about the contribution decisions of other group members. Each group member i receives an individual (private) signal s^1_{i,P_j} about the contribution of P_j ($j \neq i$) to the public good; hence A receives four signals, one for each P, and each P receives three signals, one for each of the other three Ps. The signal is given by

$$s^1_{i,P_j} = \begin{cases} e^{PG} c_j & \text{with probability } \lambda \\ e^{PG}(1 - c_j) & \text{with probability } 1 - \lambda \end{cases}$$

With probability λ the signal reflects the true underlying contribution decision of the respective peer, and with probability $1 - \lambda$ the signal states the opposite of the true underlying contribution decision of the respective peer. Since signals are private, s^1_{i,P_j} is not necessarily the same for all i; hence, while some subjects might receive a false signal about a given peer, others might receive the true signal. In this stage subjects receive new endowments, which can be used for information acquisition or punishment. Endowments are kept constant across treatments—Ps are endowed with e^{MP}, and A receives 4 e^{MP} in all treatments.

END-PRI-DEC When monitoring is endogenous and the institution is decentralized, Ps have the option to acquire further signals about the contribution decisions of other group members. P_i can acquire at most two additional signals, s^2_{i,P_j} and s^3_{i,P_j}, about the contribution decision of each P_j ($j \neq i$). Additional signals share the same properties as the first signal s^1_{i,P_j} and are independent of the previous ones. Buying a second ($b^2_{i,P_j} = 1$ if acquired, and $b^2_{i,P_j} = 0$ if not) or third ($b^3_{i,P_j} = 1$ if acquired, and $b^3_{i,P_j} = 0$ if not) signal generates costs of p^M for P_i, and for each signal that is acquired by a player P, player A also has to bear costs of p^M. P_i's profit from stage II, $\pi^{II}_{P_i}$, is given by

$$\pi^{II}_{P_i} = e^{MP} - p^M \sum_{k=2}^{3} \sum_{j \neq i}^{4} b^k_{i,P_j}.$$

A's profit from stage II, π^{II}_{A}, is given by

$$\pi^{II}_{A} = 4e^{MP} - p^M \sum_{k=2}^{3} \sum_{i=1}^{4} \sum_{j \neq i}^{4} b^k_{i,P_j}.$$

END-PRI-CEN If monitoring is endogenous and the institution is centralized, A has the option to acquire further signals about the contribution decisions of players P. A can acquire at most two additional signals, s^2_{A,P_j} and s^3_{A,P_j}, about the contribution decision of each P_j. Additional signals share the same properties as the first signal s^1_{i,P_j}. The signals acquired by A are visible to Ps as well, except for signals related to the own contribution decision. Formally, $s^2_{i,P_j} = s^2_{A,P_j}$ and $s^3_{i,P_j} = s^3_{A,P_j}$, unless $i = j$. Acquiring a further signal generates costs of $3p^M$ for A, and for each signal acquired by A, each P_i also bears costs of p^M. Let $b^2_{A,P_j} = 1$ and $b^3_{A,P_j} = 1$ if the respective signal is acquired, and $b^2_{A,P_j} = 0$ and $b^3_{A,P_j} = 0$ if the respective signal is not acquired. P_i's profit from stage II, $\pi^{II}_{P_i}$, is given by

$$\pi^{II}_{P_i} = e^{MP} - p^M \sum_{k=2}^{3} \sum_{j \neq i}^{4} b^k_{A,P_j}.$$

A's profit from stage II, π^{II}_{A}, is given by

$$\pi^{II}_{A} = e^{MP} - 3p^M \sum_{k=2}^{3} \sum_{j \neq i}^{4} b^k_{A,P_j}.$$

Stage III: Punishment

END-PRI-DEC When punishment is decentralized, Ps have the possibility to punish each other. All the signals received in stage II remain available on the screen while the punishment decisions are made. Each P_i has to decide whether to punish ($q_{i,P_j} = 1$) or not to punish ($q_{i,P_j} = 0$) other group members P_j. For each positive punishment decision, P_i has to bear costs of p^S. Similar to the case in stage II, A has to bear the costs of p^S for each punishment decision. For each P_i who punishes P_j the period profit of P_j is reduced by S. In the decentralized settings, P_i's profit from stage III, $\pi^{III}_{P_i}$, is given by

$$\pi^{III}_{P_i} = -p^S \sum_{j \neq i}^{4} q_{i,P_j} - S \sum_{j \neq i}^{4} q_{j,P_i}.$$

A's profit from stage II, π^{II}_{A}, is given by

$$\pi^{III}_{A} = -p^S \sum_{i=1}^{4} \sum_{j \neq i}^{4} q_{i,P_j}.$$

END-PRI-CEN When punishment is centralized, the whole punishment power is concentrated in the hands of A, who has the possibility to punish each P_i by choosing one of three punishment levels; the punishment choice set is given by $q_{A,P_j} \epsilon \{0, 1, 2, 3\}$. A bears costs of $q_{A,P_j} p^S$ for each P_j, and P_j's profit is reduced by $q_{A,P_j} S$. Hence, the set of potential punishment levels that Ps can receive is identical between the decentralized and the centralized punishment institutions. For example, if the authority chooses punishment level $q_{A,P_j} = 3$, then P_j receives the same punishment as P_j would have received in the decentralized treatment if all three other Ps had chosen to punish P_j. For each q_{A,P_j}, each P_i bears costs of $\frac{p^S}{3}$. In the centralized settings, P_i's profit from stage III, $\pi_{P_i}^{III}$, is given by

$$\pi_{P_i}^{III} = -\frac{p^S}{3} \sum_{j \neq i}^{4} q_{A,P_j} - S q_{A,P_i}.$$

A's profit from stage III, π_A^{III}, is given by

$$\pi_A^{III} = -p^S \sum_{i=1}^{4} q_{A,P_i}.$$

The total period profit is the sum of profits from stage I, II and III. The profit of P_i in one period is given by

$$\pi_i = \pi_i^I + \pi_i^{II} + \pi_i^{III},$$

and the profit of A is given by

$$\pi_A = \pi_A^I + \pi_A^{II} + \pi_A^{III}.$$

Table 1: Overview of Parameters

Parameter	Value	Description
e^{PG}	15 Token	Endowment for public good
e^{MP}	6 Token	Endowment for monitoring and punishment
p^M	1 Token	Price of acquiring one signal
p^S	2 Token	Price of punishment
M	2	Public good multiplier
S	8 Token	Severity of punishment
λ	0.9	Accuracy of signal

Subjects were paid the sum of period profits over all 25 periods. Period profits could be

negative, but total profits were capped at zero.[2]

Methods and Procedures

We conducted the experiments in March and June of 2015 and in November 2016 at the decision laboratory of the Department of Economics of the University of Zurich using the software z-Tree (Fischbacher, 2007). Subjects, mainly students from the University of Zurich or the Swiss Federal Institute of Technology in Zurich, were recruited using the software "hroot" (Bock et al., 2014). Subjects participated only once in this experiment. Six treatments comprise our experiment, that is, an additional four to the two treatments already described. The other four treatments are introduced in the reminder of the paper. For convenience, we describe the methods and procedures for all treatments at this points. We run a total of 16 sessions, each lasting around 90 minutes; 515 subjects participated in this study. 60 subjects participated in END-PRI-DEC, 70 in END-PRI-CEN, 130 in EXO-PRI-DEC, 125 in EXO-PRI-CEN, 70 in EXO-PUB-DEC, and 60 in EXO-PUB-CEN. Subjects were seated at computer terminals located in separate carrels. After subjects took their randomly assigned seats, they read the printed instructions and answered control questions. Subjects received an average total payment of 48.90 CHF (1 CHF $\approx$ 1.03 USD), including the show-up fee of 15 CHF.

3 Analysis

We use the cooperation rate as the key outcome variable, that is, the fraction of subjects who contributed to the public good, to analyze the impact of the underlying social norm enforcement institution in the realistic case of endogenous private monitoring. Whether decentralization or centralization fares better in such a setting is an empirical question. Previous empirical research has not brought forth a clear winner. Aspects that speak in favor of the centralized institution include the following. First, coordination of punishment, that is, if one person can make a punishment decision then there is no problem in coordinating the right amount of punishment. Assigning a certain level of punishment might prove difficult in a peer-punishment setting since peers make their decisions simultaneously. Therefore, decentralization might easily result in too little or too much punishment. Second, there is no free-riding problem on second-order public goods under centralization. The authority needs to make the information acquisition and punishment

[2]Since all subjects finished the experiment with positive total profits, we did not have to enforce the non-negative payoff restriction.

decision (both are second-order public goods), and hence cannot rely on others to enforce cooperation. Under a decentralized institution, peers have an incentive to let others buy signals and exert punishment, because these second-order public goods bear private costs, but benefit the collective. Under imperfect monitoring the potential for punishment errors threatens cooperation. Wrongful punishment of cooperators might discourage future contributions. On the other hand, the fear to wrongfully punish a cooperator might make some subjects reluctant to use the possibility to exert punishment, and therefore lets defectors to get away. Arguably, there is an advantage for the authority to keep Type-I errors lower than the decentralized institution, because under peer-punishment the probability that *at least one* peer receives a wrong signal and therefore commits a false positive is higher than that the authority receives a wrong signal (remember signals are private and drawn for each subject individually). On the contrary, under peer punishment it only requires one (out of three) subjects to receive a correct signal and to sanction a defector, whereas under centralization it has to be the authority to take actions. Due to the possibility to acquire further signals, subjects in both institutions have the means to improve their information to a point where mistakes are reasonably small. Having access to up to three signal, each one correct with 90% probability, provides a good idea about the true contribution of a subject. So if the distribution of Type-I and Type-II errors crucially depends on information acquisition and punishment behavior.

Result 1: *Under endogenous private imperfect monitoring, decentralization of social norm enforcement is superior to centralization in sustaining cooperation.*

Cooperation rates in the decentralized punishment institution are significantly higher than in the centralized punishment institution. Figure 4 shows the evolution of cooperation rates in the two treatments. Under a decentralized punishment institution the cooperation rate is on average 90.7%, which is about 7 p.p. greater than the corresponding figure of the centralized institution. The treatment difference is significant according to a Wilcoxon rank sum test based on independent average group cooperation rates aggregated over all 25 periods (p=0.0177, n=26 groups).[3] Hence, the above discussed potential

[3]We test the hypothesis $Pr(X_{treatment\ 1} > X_{treatment\ 2}) = Pr(X_{treatment\ 2} > X_{treatment\ 1})$, where $X_{treatment\ 1}$ is the outcome variable in one treatment and $X_{treatment\ 2}$ is the same variable in the other treatment. Treatment differences analyzed with the Wilcoxon rank sum test are all based on independent group averages of all 25 periods. In addition to this non-parametric tests, we also tested all treatment differences parametrically based on an OLS regression model of the form $y_{ti} = \beta_0 + \beta_1 x_i$, where x_i is a treatment dummy, and y_{it} is a dummy for the outcome variable. Standard errors in OLS regression are, if not otherwise stated, always clustered at the group level. If not otherwise stated the conclusions drawn from these analyses are qualitatively the same as those from the Wilcoxon rank sum test.

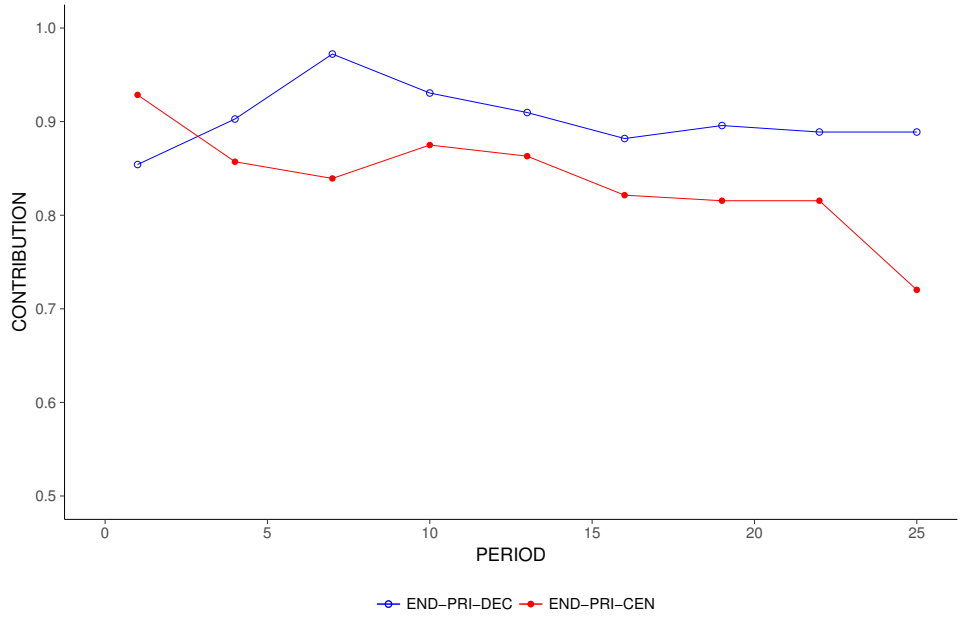


Figure 4: Cooperation Rates over Time

Note: This figure shows the fraction of subjects who decided to contribute to the public good over the 25 periods.

benefits of centralization *per se* do not match those of decentralization.

This leads to the question how the punishment institutions enforce cooperation and why the decentralized regime manages to produce higher cooperation rate. As mentioned above both types of punishment error are a threat to cooperative behavior. Naturally, not punishing defectors (Type-II errors) will miss to encourage them to start cooperating and this will in turn erode contributions of conditional cooperators. On the other hand, cooperators arguably decrease their willingness to contribute after they received punishment, that is, after a Type-I error occurred. So how does either institution fare with regard to errors of punishment?

Result 2: *There is a trade-off between Type-I and Type-II errors of punishment comparing decentralization and centralization under endogenous private monitoring. Under decentralization more cooperators are sanctioned, but much fewer defectors elude it.*

The prevalence of Type-I errors is 12.4% under decentralization, that is, in about one eighth of all cases a cooperator receives some punishment. When a single authority has all punishment power, this figure is lower by 7.1 p.p. (p=0.1211, based on an OLS regression[4]). Not only the prevalence is smaller, but cooperators receive also less severe

[4]We test for certain treatment differences by means of an OLS regression with standard errors clustered

punishment under centralization (0.54 points compared to 0.98 points), however, this difference is not statistically significant (p=0.2890, based on OLS regression).

With respect to Type-I errors, the decentralized institution performs worse than its centralized counterpart. Notwithstanding, decentralized institution has a decisive advantage regarding Type-II errors. In END-PRI-CEN more than 43% of all defectors go unpunished, which is about 2.3 times as many as under peer punishment (18.8%), this difference is significant (p=0.0137, based on Probit regression). The received punishment by defectors is only slightly higher though (9.58 points compared to 11.14, p=0.5512 based on OLS regression). Figures 6 and 7 depict the average prevalence of Type-I and Type-II errors for all (including the not yet introduced) treatments. This pattern suggests that decentralization manges to sustain higher cooperation rates by punishing more defectors and that the negative effects of more wrongfully punished cooperators is not enough to offset this advantage.

Note that subjects can always completely avoid one of the errors at the cost of maximizing the other. To never punish will result in zero Type-I errors, but 100% Type-II errors. Always punishing would result in zero Type-II error, but 100% Type-I error. The error structure is a combination between monitoring structure, institution, information acquisition behavior and punishment decisions. Assume subjects want to punish defectors and to spare cooperators. Then a subjects who focuses on the reduction of Type-I errors should buy additional signals when the initial signal states "Defector" in order to make sure that it is indeed a defector. Focusing on Type-II error rates would mean to acquire new signals when the initial information depicts a group member as a "Cooperator". What is subjects' behavior regarding information acquisition, do they even make use of this costly possibility and if so do they focus on avoiding Type-I or Type-II errors?

Result 3: *There is substantial demand for additional information about the actions of other group members who appear as defectors (focusing on avoiding Type-I errors), both under decentralization and centralization.*

at the group level. Non-parametric Wilcoxon rank sum tests are in the case of error rates (Type-I and Type-II) not appropriate, because the base rate of cooperators and defectors are not equal across groups. This means that using a single a single aggregate observation per group would overweight groups that consists of only a few of a certain type. For instance, a group in which only one subject once does not contribute to the project and goes unpunished would generate a Type-II error rate of 100% (based on a single observation). The error rate of a group with relatively many defectors would receive the same weight as the first one. When we compare treatment differences based on an OLS regression we use the form $y_{ti} = \beta_0 + \beta_1 x_i$, where x_i is a treatment dummy, and y_{it} is a dummy for the outcome variable, for instance, "Type-I error" of subject i in period t. Only observations that are relevant for the respective treatment difference are included in the analysis (e.g. for Type-I error rate only cases where the subject actually contributed).

Figure 5 shows the fraction of cases in which a second signal is acquired in the endogenous monitoring treatments, depending on the first signal. While a second signal is acquired in over 50% of the cases in which the first signal states "Defector", a second signal is acquired in less than 10% of cases in which the first signal states "Cooperator". Signal acquisition behavior does not differ between the decentralized and centralized settings, neither after a first signal "Defector" (p=0.6747 based on OLS) nor after a first signal "Cooperator" (p=0.9025 based on OLS). While a third signal is acquired in only 5.9% of the cases in which the first two signals state "Cooperator," a third signal is acquired in 27% of the cases in which the first two signals state "Defector" (p=0.0050 based on OLS regression). If the first two pieces of information reveal conflicting signals, a third signal is acquired in 36.6% of the cases.

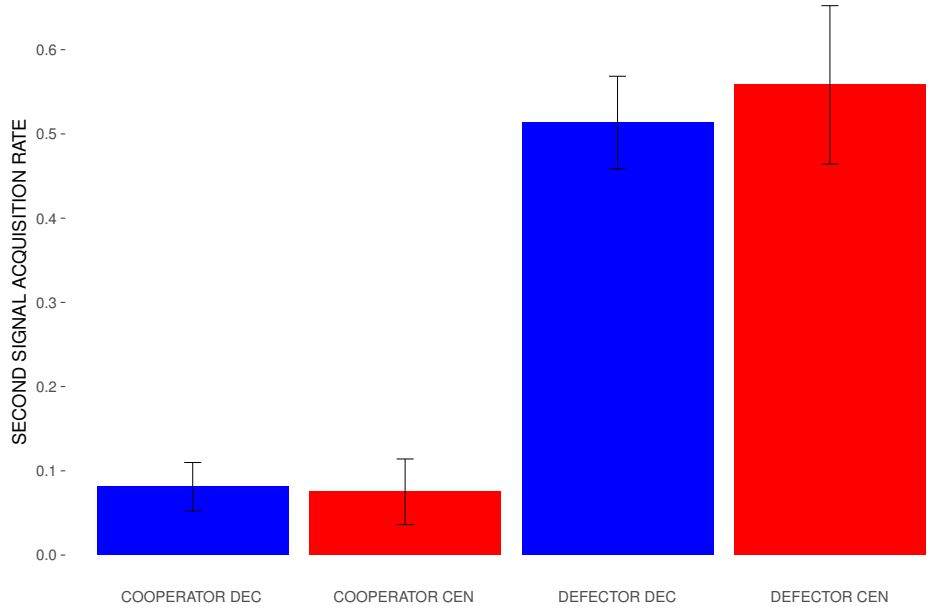


Figure 5: Acquisition of Second Signal Depending on First Signal

Note: This figure shows the fraction of cases in which subjects acquire a second signal in the endogenous monitoring treatments (i.e. END-PRI-DEC and END-PRI-CEN) depending on the first signal. Bars indicate clustered standard errors.

Subjects seem to focus on avoiding Type-I errors of punishment by acquiring further information when the initial one states "Defector," but does this indeed diminish the prevalence of false positives? How acquired signals translate into a change in punishment error patterns is not a trivial question, because these patterns emerge as a combination of several factors as outlined above. In order to answer this question we run an additional set of experiments, where the subjects do not have the possibility to buy new signals, that is, information is exogenous.

EXO-PRI-DEC is the same as END-PRI-DEC with the exception that peers do not have the possibility to acquire new signals in stage II monitoring. Subjects just receive a single signal and peers have to base their punishment decision on this one piece of information (exogenous information). The total endowment of peers for stage II and stage III remains unchanged, that is, in both treatments peers have the same funds to finance second-order public goods (in END-PRI-DEC information and punishment and in EXO-PRI-DEC only punishment).

EXO-PRI-CEN is the same as END-PRI-CEN with the exception that the authority does not have the possibility to acquire new signals in stage II monitoring. Subjects just receive a single signal and the authority has to base its punishment decision on this one piece of information (exogenous information). The total endowment of the authority for stage II and stage III remains unchanged, that is, in both treatments the authority has the same funds to finance second-order public goods.

Result 4: *Subjects employ costly information acquisition to reduce Type-I errors. Type-II errors on the other hand remain largely unaffected. The trade-off between Type-I and Type-II errors is also present under exogenous private monitoring when comparing decentralization and centralization.*

Type-I error rates are indeed larger, when subjects' information was exogenously given. In EXO-PRI-DEC, Type-I errors are almost 2.5 times as large as in END-PRI-DEC (p=0.0027, based on OLS). At the same time Type-II errors are virtually unaffected (-0.4 p.p., p=0.9604 based on OLS). Also, in EXO-PRI-CEN Type-I errors are more likely and occur in 16.7% of all cases, which is significantly (p=0.0242 based on OLS) more often than in END-PRI-CEN. In EXO-PRI-CEN Type-II errors are also larger by 12.5 p.p. compared to END-PRI-CEN, but this difference is not significant (p=0.1463 based on OLS). Unlike Type-II error rates, Type-I error rates clearly improve under endogenous private monitoring, because subjects focus on the acquisition of additional signals of group members who appear as defectors. Analogously to endogenous monitoring, under exogenous private monitoring the decentralized institution lets fewer defectors escape punishment at the cost of a higher prevalence of punished cooperators.

So subjects make use of the endogenous monitoring possibility and mange to reduce the probability that a cooperator receives punishment. This is certainly a normatively desirable effect, however, this does not imply that there is also a positive effect on cooperation rates. Subjects in both monitoring settings have a single account to buy new signals and

to finance sanctions, that means, resources spent on more information may be lacking to actually punish norm violators. Furthermore, as we outlined above, the problem with smaller cooperation rates seems to be mainly due to the lack of punished defectors and not a problem of too many wrongfully punished cooperators. Nevertheless, subjects use the information acquisition opportunity in a way that not only reduces Type-I errors, but also boosts contributions to the public good.

Result 5: *Subjects in both institutions mange to use the possibility to endogenously improve information to foster cooperative behavior.*

Without the possibility to improve information the cooperation rate under peer-punishment is more than 10 p.p. lower, which is a significant difference to the setting with endogenous monitoring (p=0.0135 based on Wilcoxon rank sum test, n=38 groups). The centralized institution faces an even greater decrease in cooperation of almost 16 p.p. when subjects cannot acquire further information (p=0.0095 based on Wilcoxon rank sum test, n=39 groups). The conclusion about the relative performance of institutions remains the same under exogenous private monitoring as before, although Type-I errors are of greater concern in decentralization, and one therefore might expect peer punishment to lose its edge when peers cannot gather further information to reduce those errors.

Result 6: *Under exogenous imperfect private monitoring, decentralization is superior in sustaining cooperation compared to centralization.*

Cooperation rates in EXO-PRI-DEC are on average 80.1% compared to to only 67.3% in EXO-PRI-CEN (p=0.0307 based on Wilcoxon rank sum test, n=51 groups). Figure 8 depicts the average cooperation rates of all treatments.

Up to this point all treatments feature *private* imperfect monitoring. One might argue that the superiority of decentralization vanishes when information is made public instead of private. The reasoning behind this argument is that under decentralization there seems to be more, or even more accurate, information, because those who hold punishment power have access to more signals. The authority has four signals (one per peer) and all peers combined have twelve signals (three per peer). The issue with that line of reasoning is, however, that only aggregate information is more accurate. Despite the more accurate aggregate information, there is also a higher chance of some false information of a punisher when signals are private. The trade-off between the institutions

regarding the error structure already indicates that decentralization does not mange to use its alleged information advantage to keep both error rate below those of its centralized counterpart. Whether or not private instead of public signals constitutes an advantage for the decentralized punishment institutions is an empirical question. Two additional treatments are conducted to investigate whether private signals are a blessing or a curse for either institution. We expect the centralized institution to be basically unaffected by making signals public instead of private, since there is no change in the information structure of the authority who holds all sanctioning power. But the perfect correlation of signals received by peers potentially changes, for instance due to conditional cooperation, how contributions evolve. We expect this perfect correlation to impact the decentralized institution stronger since those with punishment power, the peers, now all have the same signals. As pointed out above, changes in contributions could go in either direction. If every peer has the same signals, then Type-I errors possibly become less prevalent, Type-II errors might increase, because the likelihood that all peers receive a wrong signal is now exactly 90%, whereas before three independent signals had to be false at the same time.

EXO-PUB-DEC is the same as EXO-PRI-DEC with the exception that all signals are public instead of private. Formally, signals about peer j are the same for all group members, that is, $s^1_{i,P_j} = s^1_{k,P_j} = s^1_{k,A}$.

EXO-PUB-CEN is the same as EXO-PRI-CEN with the exception that all signals are public instead of private. Formally, signals about peer j are the same for all group members, that is, $s^1_{i,P_j} = s^1_{k,P_j} = s^1_{k,A}$.

We find the following; in contrast to the alleged information advantage a decentralized institution gains from private signals,

Result 7: *Decentralization profits from public imperfect information compared to private imperfect information, and produces higher cooperation rates and lower Type-I error rates. Centralization is unaffected by public imperfect monitoring compared to private imperfect monitoring. Decentralization is also in this setting superior in sustaining cooperation.*

Cooperation rates in EXO-PUB-DEC are on average 88.4%, which is significantly (p=0.0210 based on Wilcoxon rank sum test, n=40 groups) more than when information is private

instead of public. By making information public Type-I error rates decline by 15.6 p.p. (p=0.0000 based on OLS) under decentralization. There is almost no effect on Type-II errors (-1.1 p.p., p=0.8388 based on OLS). The centralized institution does not profit from making information public. Cooperation rates in EXO-PUB-CEN are on average 67.3%, which is an insignificant 1.8 p.p (p=0.9482 based on Wilcoxon rank sum test, n=26 groups) below those in EXO-PRI-CEN. Error rates are similar in EXO-PUB-CEN and EXO-PRI-CEN. Cooperators receive punishment in about 10.5% of cases (+1.00 p.p., p=0.7798 based on an OLS) and defectors elude it in 55.8% of instances (+3.11 p.p., p=0.7537 based on OLS). Finally, decentralization is also under this monitoring structure superior in sustaining high cooperation rates as it was already the case with both other information environments (p=0.0033 based on Wilcoxon rank sum test, n=26 groups).

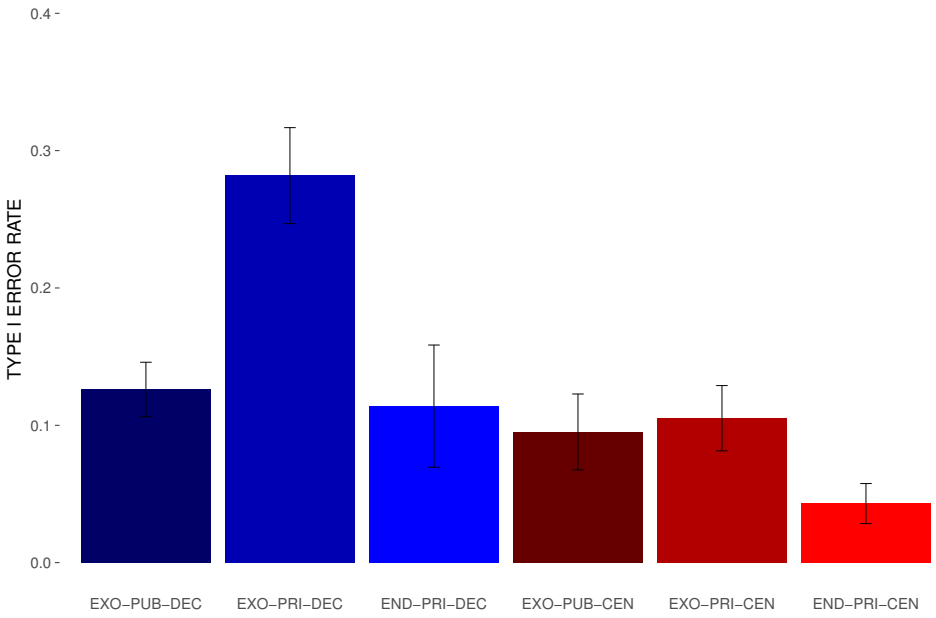


Figure 6: Type-I punishment error rates for all treatments

Note: This figure shows the fraction of cooperators who received punishment (Type-I error). Error bars denote clustered standard errors.

4 Conclusion

In this paper, we study human behavior and the relative performance of social norm enforcement institutions under monitoring technologies that exhibit more realistic features than generally assumed in the literature. In a laboratory public goods game, we systematically vary whether imperfect signals about contributions are exogenous public,

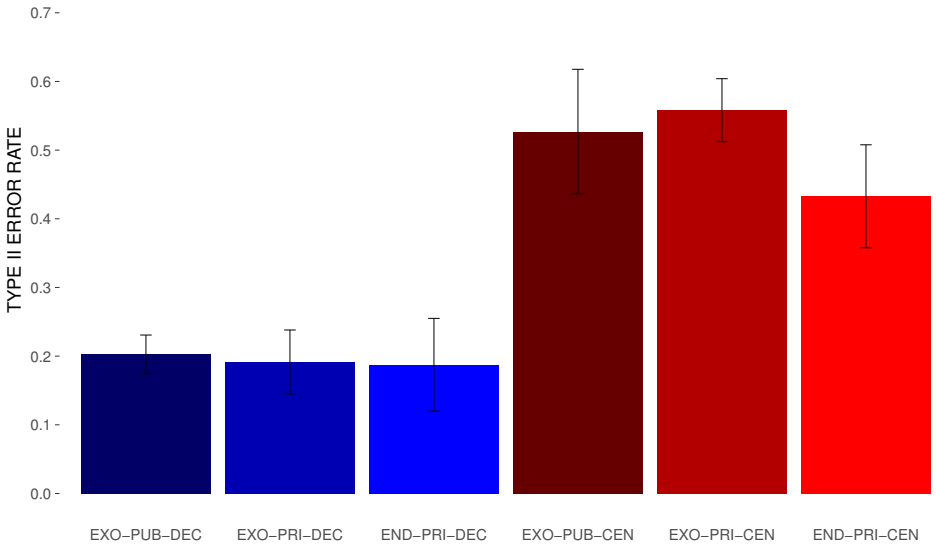


Figure 7: Type-II punishment error rates for all treatments

Note: This figure shows the fraction of cooperators who received punishment (Type-II error). Error bars denote clustered standard errors.

exogenous private or endogenous private. We compare a decentralized peer-to-peer punishment institution with a centralized setting, where all punishment power is delegated to a randomly selected authority.

The results show that under all considered information structures decentralization achieves significantly higher cooperation rates than centralization, because defectors are more likely to receive sanctions when punishment is decentralized. The two punishment institutions involve a trade-off between lower Type-II punishment error rates (not punishing a defector) in the decentralized setting and lower Type-I punishment error rates (punishing a cooperator) in the centralized setting, but the benefits of the lower probability that defectors remain unpunished in the decentralized punishment institution outweigh the disadvantages of the higher likelihood of sanctioning cooperators.

Moreover, we find substantial demand for additional signals about the contribution decisions of other group members. In the endogenous monitoring treatments, where subjects can acquire information in addition to the initial signals, subjects are willing to incur costs to improve their information base before exerting punishment; in particular, subjects focus their information acquisition on group members who appear as defectors. By establishing a "standard of proof" before exerting punishment subjects significantly reduce Type-I errors of punishment under endogenous monitoring, boosting cooperation rates compared to an environment where costly information acquisition is not possible.

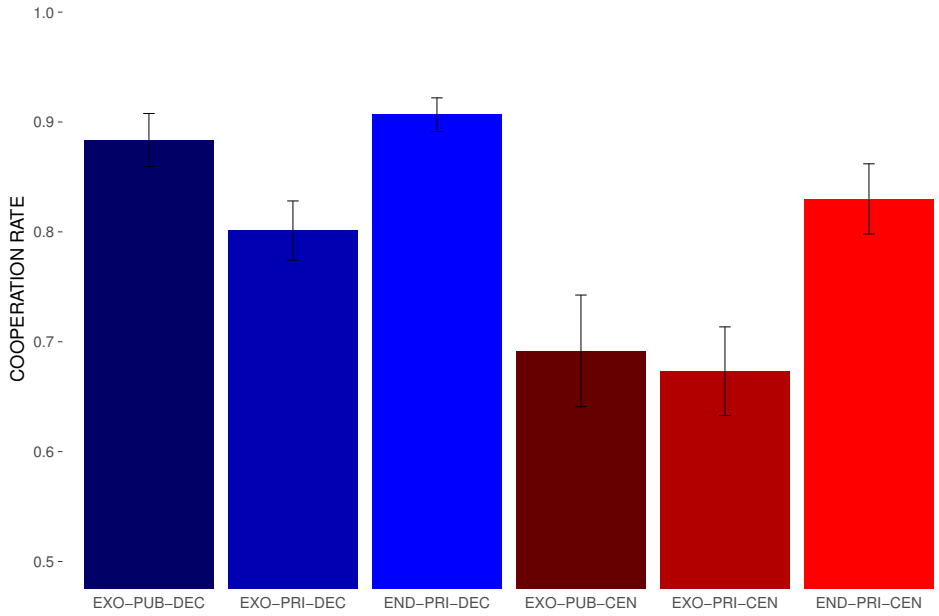


Figure 8: Cooperation rates all treatments

Note: This figure shows the fraction of cooperators. Error bars denote clustered standard errors.

Furthermore, we show that private signals, which means having access to more information in the aggregate, is a curse rather than a blessing for a decentralized sanction institution. Having a perfect correlation between signals (public signals) leads to more contributions than when signals are private. Centralization is unaffected by this change in the monitoring technology.

The findings of this paper suggest that under imperfect monitoring centralization of social norm enforcement per se leads to inferior resolutions of social dilemmas compared to a peer-to-peer punishment environment and that centralization needs to be associated with other performance-enhancing features—such as commitment to punishment rules or election of authorities—to become as effective as peer-to-peer punishment institutions in solving problems of collective action.

Chapter IV

The Dynamics of Norm Formation and Norm Decay

The Dynamics of Norm Formation and Norm Decay

Ernst Fehr & Ivo Schurtenberger

Abstract

Social norms are an ubiquitous feature of social life, and pervade almost every aspect of human social interaction. However, despite their importance we still have relatively little empirical knowledge about the forces that drive the formation, the maintenance and the decay of social norms. In addition, due to the lack of exogenous variation in norms, knowledge about their causal effects is also limited. We tackle these questions with the help of a laboratory public goods experiment in which we allow subjects to explicitly formulate normative requests about the contributions that every group member should make. This approach enables the empirical examination of the dynamics of social norm formation and norm decay. When decentralized private enforcement of norms is possible, strong, stable and demanding social norms emerge, which are also largely obeyed. In stark contrast, when punishment of norm violators is ruled out, not only weaker, less stable and less demanding norms arise, but these norms are also regularly disobeyed. In addition, the opportunity to formulate normative requests unambiguously causes higher public good contributions and group welfare, but only under peer-punishment. The norm formation opportunity renders peer-punishment more efficient than a no punishment control, which is not the case without it. Furthermore, without punishment, normative requests are merely cheap talk and do not exhibit any positive effects.

JEL classification: C92; A13; H41

Keywords: Public goods; Social norms; Norm enforcement; Norm formation; Norm decay; Cooperation

Citation: Fehr, E., & Schurtenberger, I. (2018). The dynamics of norm formation and norm decay. *Working Paper*.

1 Introduction

Social norms are the fabric of human social interaction. They are omnipresent, and permeate almost every aspects, from mundane to profound, of social life. These include, for instance conformity (Bernheim, 1994), involuntary unemployment (Akerlof, 1980), tipping (Conlin et al., 2003), littering (Cialdini et al., 1991), cooperation (e.g. Ostrom, 1998) and altruism (Krupka & Weber, 2013). Akerlof (2007) even argues that the five neutralities of neo-classical macroeconomics do not hold when norms are reasonably incorporated into the framework.

Despite their importance, there is, however, a lack of empirical knowledge about two of the major questions regarding social norms. First, what are the underlying forces that drive the formation, the maintenance and the decay of social norms? Second, what is the causal effect of a social norm on human behavior?

Our novel approach permits us to deepen our understanding of both questions. Even though one can reasonably hypothesize that the sword is no prerequisite for the obedience of social norms (e.g. Bicchieri, 2006; Ostrom et al., 1992), we do not only find the opposite to be the case, but more astoundingly that even the formation and decay of the social norm itself to crucially depend on means of enforcement. Regarding the second major question, we reveal social norms to exhibit a positive causal effect on behavior, ranging from cooperative behavior over sanctioning patterns to one's reaction to punishment, when norm violators can be sanctioned, but to be merely cheap talk, without any positive influence on behavior, when sanctions are ruled out. Our results are based on a controlled laboratory public goods experiment that allows to empirically assess social norms and to exogenously manipulate social norms under various enforcement institutions.

The dynamics of social norm formation and their causal effect on behavior pose important, yet difficult to solve, challenges. First, norms are inherently difficult to quantify, since they are sophisticated entities featuring multiple properties. These properties include the prescription itself, but also to what degree the social norm is accepted and how prone properties are to change over time. Second, social norms are based on social dynamic processes that are, by virtue, influenced by many interdependent variables that are often not subject to exogenous variation. This renders the establishment of causal relationships between these variables and the formation of norms and their effects on behavior tough. Consider, for example, the availability of punishment opportunities in the case of contribution norms in a team production setting. A team that struggles with the formation and subsequent obedience of strong social norms may introduce a punishment

mechanisms to cope with the problem. A team without such issues, on the other hand, has no need for, and may therefore lack a sanctioning scheme. Comparing these two teams would then suggest that punishment opportunities are generally unnecessary or even harmful for the formation of social norms and their obedience. Third, the dynamics of norm formation normally takes place over longer, and often disruptive, periods of time, which makes their evaluation often impractical. Finally, the lack of credible exogenous variation in social norms renders the establishment of causal effects on behavior difficult.

We tackle the aforementioned difficulties by introducing a simple feature into a well established experimental design. We provide subjects with a *norm formation opportunity*. This method enables the exogenous variation of the social norm in order to establish a causal relationship between prescription and action, and at the same time allows us to quantify key properties of social norms. The laboratory provides a controlled environment to exogenously manipulate variables that are potentially important to the formation and causal effect of social norms. Additionally, this controlled environment enables the observation of dynamics that probably would require an unrealistic amount of time in a field setting.

We understand a social norm as standards of behavior that are based on widely shared beliefs how individual group members ought to behave in a given situation (Fehr & Fischbacher, 2004a). In that case, what are the key properties of a social norm? We recognize three such key properties, namely (a) their content, (b) their strength and (c) their stability. Naturally, a social norm needs a content, that is, a prescription of actions that is regarded as obligatory, permitted, or forbidden (Crawford & Ostrom, 1995; Ostrom, 2000). In case of a cooperation norm, the content of a social norm is characterized by how demanding it is to adhere, that is, demanding norms require the individuals to incur high costs. Less demanding norms on the other hand ask individuals to take actions that are less costly to them. We understand the strength of a social norm as the consensus about the appropriateness of an action, that is, to what degree the members of a community agree on, or in other words share, the aforementioned content as normatively appropriate. This means that strong social norms exist when there is little or no disagreement about appropriate behavior among the members of a group. On the contrary, the content of a weak social norm is not widely recognized As social norms are not set in stone, they are prone to change for better or for worse. By the stability of a social norm we mean how rigid its content and strength are.

Specifically, a repeated public goods game serves as a paradigm that holds the potential for the formation of a cooperation norm. We choose a public goods experiment for

the following reasons. First, the setting needs to exhibit room for the formation of a social norm. Researchers (e.g. Elster, 1989b; Ostrom, 1998; Fehr & Gächter, 2000b) have argued social norms to play a decisive role in such problems of collective action. Second, the potential norms should fit a rather simple description, such as "contribute X to public good," in order to quantify them. And third, the empirical knowledge of the norm formation and their causal effects should provide direct insights into an important area that potentially benefits a great deal from social norms. Humanity's progress and prosperity decisively depend on its members' willingness to cooperate (e.g. Axelrod, 1980; Dawes, 1980). But, establishing and sustaining cooperative behavior is neither trivial nor assured since the interest of a group as a whole and the one of its members are often unaligned. The pursuit of one's own best self-interest, therefore, produces oftentimes socially undesirable outcomes. Examples range from team production to national defense spending.

In this setting we offer subjects an opportunity to form a social norm about appropriate actions, that is, about how much each group member should contribute to a common goal. Each period subjects are asked to indicate how much each group member *should* contribute to the public good. We then merely ensure that there is a (endogenous) content, by conveying the average of requested contributions to the whole group. The majority's opinion is arguably the best candidate for the content of a social norm in our set-up. However, the formation of a social norm further requires that this content is a shared understanding about appropriate actions. Our mechanism allows measuring the strength of a social norm, because we receive each group member's individual assessment of how much should be contributed, and therefore, allows estimating a proxy of the degree of agreement. The closer the answers of subjects are, the stronger the social norm. Repeating the interactions reveals how the content and the strength evolve over time depending on previous behavior, thus allowing us to analyze the dynamics of norm formation. We exogenously vary the punishment institution in order to establish a causal effect of means of norm enforcement on social norms (e.g. Fehr & Gächter, 2000a). The employed punishment institution is a relatively weak one, since subjects have the option to retaliate against their punisher in form of counter-punishment (e.g. Denant-Boemont et al., 2007; Nikiforakis, 2008). We include this feature, since we want to examine whether the formation and effects of social norms depend on enforcement, even when enforcement is only possible in a clearly suboptimal manner. Furthermore, this allows us to examine the causal effects of social norms on retaliation.

How are social norms expected to influence behavior our experiment? There are several potential channels that can become active regardless of punishment possibilities. First,

group members might be intrinsically motivated to follow social norms, and therefore, increase their contributions to the public good when such a norm is adopted by the group, respectively the guilt one would feel by violating the norm directly causes conformity with it (Elster, 1989a). Second, formed social norms may serve as signals for higher average contributions, which would lead conditional cooperators (e.g. Fischbacher et al., 2001) to contribute to a larger extent. Third, in her seminal work Bicchieri (2006) argues that one of the reasons to obey a social norm is "that one accepts others' normative expectations as well founded." A violation of the social norm would require the group member to justify, even if only to himself, his deviation by offering better reasons than those of the other group members. When sanctioning is an option, contributions are expected to be greater if formed social norms increase the credibility of punishment in case of free-riding. Finally, social norms potentially shape the reaction to received punishment. Punished free-riders, now branded as norm violators, may react more strongly, and change their behavior more drastically.

Thus, social norms may or may not require the sword to become operative. Notwithstanding, social norms need to be present in either case to shape behavior. So, how may punishment opportunities drive the formation of the prevailing group norm itself? There is per se no reason to assume that a sanctioning institution is required to form strong, stable and demanding norms. "Everyone should contribute such that the social surplus is maximized" is, for instance, a normatively appealing request, no matter whether this norm can be enforced subsequently. This norm might very well persists even if group members should not honor it with their actions. Bicchieri (2006, p. 11) explicitly states that sanctions may or may not be a condition for a social norm to exist. Furthermore, she outlines that social norms can in fact exists despite not being followed (p. 27). Hence, even when sanctions should be necessary to prevent the breakdown of contributions, this would not imply that the social norm itself has to decay with it.

We find that strong and stable social norms, demanding contributions close to the surplus maximizing level, emerge, but only when punishment is possible. These norms are in turn largely obeyed by subjects. In stark contrast, when peers cannot sanction each other, there is not only substantial disagreement about the appropriate behavior, but the content is less demanding as well. This difference in the social norm only appears over time, that is, subjects show similar expectations about appropriate behavior at the beginning of the experiment regardless of the availability of punishment. Moreover, subjects regularly violate the prescribed actions and make far smaller contributions to the public good than demanded without the threat of punishment. Taken together, we conclude that without means of enforcement, norms of cooperation quickly decay, whereas strong norms of

cooperation are formed and sustained when peers have the power to sanction each other, even though the environment is hostile to enforcement due to the presence of counter-punishment.

The dynamics of norm formation and norm decay are confirmed by a second set of experiments assessing the social norms with the Krupka-Weber method (Krupka & Weber, 2013). We let subjects evaluate the social appropriateness of different contribution levels to the public good in several scenarios. The first scenario is the first period when punishment opportunities exist. The second scenario entails also the first period, but when punishment is not possible. In either scenario the social norm is clear: high contributions to the public good. The third scenario describes the last period of a setting with punishment. Groups always requested high contributions and subsequently obeyed this request. In this case, the social norm solidifies, that is, high contributions become even more appropriate and medium and low contributions become even less so. The fourth scenario describes the last period of group that had no punishment opportunity. Requests declined over time and disobedience was present over the course of the first 14 periods. In this instance, the social norm clearly decays compared to the first period assessment. High contributions are now less socially appropriate than medium contributions, and low contributions are much less condemnable. Taken together, consistent demanding requests and their obedience—as it is regularly the case with punishment—foster social norms of highly cooperative behavior. Declining and disobeyed requests—the prevailing pattern in social dilemmas without punishment—lead to a decay in the social norm itself.

The experiment further reveals that social norms cause greater contributions and group welfare only when norms can be enforced. This interaction between norms and enforcement is present even though sanctions faces the imminent threat of retaliation. The content of the norm proves to be a powerful predictor of actual contributions, even when controlling for established predictors. The magnitude of this norm effect is similar to the one of conditional cooperation. Otherwise, without punishment, we do not observe any positive effect on contributions nor on group welfare. Hence, whether or not social norms are merely cheap talk decisively depends on the ability to enforce them. Of particular interest are the positive causal effect of social norms on group welfare under peer punishment. Gains from higher contributions need not be realized in final payoffs, as demonstrated by several experiments (e.g. Fehr & Gächter, 2002; Gürerk et al., 2006; Herrmann et al., 2008; Gächter et al., 2008) comparing peer punishment with no punishment, due to the social costs of punishment. Norms could very well increase the resources spent for and devoured by punishment by calling for increased severity. But on the contrary, the formation of a cooperation norm not only increases group welfare under peer

punishment, but also render peer punishment more efficient than no punishment, which is not the case without the opportunity to form a norm. This suggests that the feature of norm formation should be incorporated when evaluating the efficacy of punishment institutions. The results of Fehr & Williams (2018) also support this suggestion. In their study, peer punishment without the opportunity to form a norm does not emerge endogenously when there is the alternative of peer punishment, but with a norm formation opportunity. We identify the following reason for the striking difference in the relative advantage of peer punishment when norm formation is enabled instead of ruled out by experimenters. Free-riders react more strongly to received punishment under norm formation. This manifests in sustaining higher cooperation rates while punishing free-riders less severely.

Additionally, we observe the norm coordination opportunity to affect punishment behavior in three ways. First, significantly less punishment of free-riders was exerted without negatively affecting contributions. Secondly, there is evidence that norms decrease antisocial punishment, arguably the clearly stated average expectation renders the punishment of above average contributions less legitimate. Without norm formation, there is significant punishment of above average contributors. Thirdly, groups increase their punishment severity the higher average contributions are, and therefore push for the really high levels when norms can be formed. Finally, we obtain results on counter-punishment behavior. The more someone deviated negatively from average contributions of others the less severely this subject retaliates against its punisher, an effect that is slightly more pronounced with the norm formation opportunity. Under such explicit social norms, above average contributors are also significantly more likely to engage in retaliation, possibly since they regard the received punishment as less legitimate and therefore feel encouraged to defend themselves against uncalled punishment. Our data suggests that counter-punishment is mainly driven by reciprocity concerns and not by strategic considerations to deter future punishment. Such a strategic use of counter-punishment would likely prove unfruitful anyway, since subjects do not seem to decrease their exerted punishment due to previously received counter-punishment.

Our results talk to several strands of the literature. We are the first to explicitly quantify the dynamics of social norms in a public goods setting. This contributes to the large body of studies that argued that social norms are decisive for either the maintenance or breakdown of cooperation (e.g. Ostrom, 1998, 2000; Elster, 1989b; Fehr & Gächter, 2000a). This allows for a novel interpretation and a deeper understanding of the mechanism at work in studies concerned with collective action problems. Consider a regular public goods game, it was impossible to tell apart if declining contributions are due to

an increase in norm violation alone without a change in the norm itself, or whether the social norm itself decays in such a scenario as well. Our data suggests that the latter constitutes a substantial part of the whole picture.

Furthermore, our research is related to the extensive literature about how mankind has overcome collective action problems in so many instances, contrary to the worrying prediction of classical theory and the dire results from lab-experiments of standard public goods without punishment or communication. An early contender for such a remedy was communication. Verbal face-to-face communication has proven to be an effective mean to sustain very high levels of cooperation (e.g. Isaac & Walker (1988); Ostrom et al. (1992) or for an early overview of 36 studies Sally (1995)). However, Bochet et al. (2006) show that numerically communicating one's "contribution intention" does not raise actual contributions neither with nor without punishment. Our norm formation opportunity is also purely numerical, but increases contributions when sanctions are possible. Hence, verbal communication is not required, but the efficacy of numerical communication seems to be much more effective when it provides the opportunity to form social norms at least when violators can be sanctioned.

We also contribute to the literature of the influence of social norms on behavior. Recent research has made progress in that regard by showing that norms can be used to make predictions about behavior instead of 'just' using them as a post-hoc interpretation of observed phenomena (Krupka & Weber, 2013). However, empirical evidence for the causal relationship between a norm and behavior is rare. Our data reveals that the effect of an equally pronounced norm, with respect to content and consensus, crucially depends on the availability of means of enforcement. By relying on researchers' interpretation of a norm, that is, without the possibility to quantify it, it is difficult to prove that a social norm influences behavior in one setting and not in another, since the norms might differ in the settings.

Finally, this paper is related to the literature on counter-punishment. Previous studies report the breakdown of cooperation under counter-punishment in very similar settings to ours. The "revenge only" treatment in Denant-Boemont et al. (2007) and "PCP" treatment in Nikiforakis (2008) both show such a pattern. In contrast, contributions in our samples from two different universities do not break down even without the norm formation opportunity. An unanswered question therefore seems to be under what conditions counter-punishment poses a serious threat to the efficacy of peer-sanctioning. Nikiforakis et al. (2012) show that feuds, a sequence of counter-punishment, is especially threatening to high contributions under normative conflicts, that is, when not everyone benefits to

the same degree from contributions.[5] The authors themselves note that enabling subjects to resolve their normative conflict might mitigate the problem of feuds. Our results on counter-punishment and the effect of the norm formation opportunity point indeed in this direction.

The results of this study might prove useful to shape policies that aim at improving cooperative behavior. There is a call for such policies in many areas, including littering in public parks (Cialdini et al., 1991), improving teamwork between physicians and nurses (Makary et al., 2006), improving tax compliance (Andreoni et al., 1998), encouraging pro-environmental behavior (Steg & Vlek, 2009) or rebuilding and managing fishery in a sustainable way (Botsford et al., 1997; Worm et al., 2009). We believe that incorporating norm formation opportunities in these situations to be worthwhile.[6]

The rest of the paper is outlined as follows. Next, we describe the experimental design and the procedure of our study, then we proceed with our empirical findings and finally we conclude.

2 Experimental Design

In a laboratory experiment we offer subjects a simple device to form explicit social norms about appropriate behavior in a social dilemma. The paradigm of a public goods game poses an exemplary situation in which there is a stark contrast between the interest of a group as a whole on one side and the individuals that compoase the group on the other side. Pareto improvements can be obtained when subjects obey social norms commanding them not to act in an exclusively selfish manner. A distinctive characteristic of such a setting is that social norms and behavior are readily quantifiable in form of the extent of contributions, that is, social norms take the simple form of "everyone should contribute X."[7] This allows an empirical investigation of social norms.

All treatments are built around a linear public goods game, which is repeated for 15

[5]The authors argue that two normative points of view compete with each other in this setting. The norm can either take the form "Everyone should *contribute* X" or "Everyone should *earn* Y". If endowments and the MPCR are the same for all subjects than these two contenders for a social norm prescribe the same, else they do not.

[6]For this, there is further, more anecdotal, evidence from a consulting project in which one of the authors was involved. A large Austrian media house struggled by the lack of collaboration between several of its business units. None of the policy proposed by different large international consulting companies bore fruit and the situation remained dire. Until the heads of these business units were brought together and were encouraged to explicitly form themselves norms about what they expect from one another.

[7]There are no asymmetries in endowments nor in benefits from the project, hence it is only natural that everyone should behave the same in this situation.

periods. Four randomly selected subjects form a group. Groups remain together for the whole experiment, but the identification number of subjects changes from one period to the next to avoid reputation effects and mitigate the problem of spillovers, for instance, in punishment, across periods. Depending on the treatment every period consists of up to four stages: Norm Formation (stage 1), Contribution (stage 2), Punishment (stage 3) and Counter-Punishment (stage 4). The stages, when present in the treatment, are basically identical in all treatments. We employ a 2x2 factorial design by varying both the opportunity to form explicit social norms and the availability of punishment as well as counter-punishment possibilities. Table 2 lists all treatments and their corresponding label. The richest and arguably most realistic environment provides treatment NF. In this treatment subjects go through all stages, that is, they first submit normative requests and receive the average answer of the group, second, they make their decision about how much to contribute to the public good, third they are informed about the contributions of others and have the possibility to costly assign reduction points to lower others' income, and finally they are given the option to retaliate against their punishers by reducing their payoff.

Treatment NFnoP enables studying the implications of means of (an absence of) norm enforcement on the emergence and on the behavioral influence of social norms. The first two stages in NFnoP are the same as in NF, but no punishment and naturally no counter-punishment opportunities exist. NF and NFnoP are the two treatments with an explicit norm formation opportunity. We include two further treatments, noNF and noNFnoP, in order to study the causal effect of social norms on the level of cooperative behavior and group welfare as well as the punishment and counter-punishment behavior. Treatment noNF is the same as NF, but without explicit norm formation. Analogously noNFnoP corresponds to NFnoP without norm formation, hence, a standard public goods game. In the following we describe the stages in more detail.

	NF	**noNF**	**NFnoP**	**noNFnoP**
Norm Formation	YES	NO	YES	NO
Contribution	YES	YES	YES	YES
Punishment	YES	YES	NO	NO
Counter-Punishment	YES	YES	NO	NO

Table 2: Overview of Treatments

Stage 1 Norm Formation Opportunity

Subjects are asked *In your opinion, how many Token should each group member contribute to the project?* They have to answer this question with an integer between 0 and 20, which cover all possible contribution levels. The content of a social norm is then conveyed to all group members in form of the statement *According to the average opinion of your group each group member should contribute the following number of Token:* followed by the mean of answers (rounded). Note that the question clearly asks subjects to indicate what behavior they regard as appropriate, by asking about subjects' opinion about what *should* be done. Furthermore, the way the mean is conveyed closely resembles key characteristics of social norms, namely that the request is regarded as appropriate by an average group member and that the request itself originates from the opinions of group members themselves. The average opinion about contributions remains visible at the top of subjects computer screen throughout all further stages.

Stage 2 Contribution to Public Good

In the contribution stage, which is the only one present in all treatments, each subject i receives an endowment of $e^{PG} = 20$ Token, and decides the amount $c_i \epsilon \{0, 1, 2, ..., 20\}$ she wants to contribute to the public good[8]. This gives us a broad enough range of possible contributions and explicit norms, and keeping the decision space simple at the same time. Contributions are multiplied by a factor of 1.6 and redistributed evenly across all group members, that is, the marginal per capita return is 0.4.[9] Monetary payoff after stage 2 is given by

$$\pi_i^{II} = e^{PG} - c_i + \frac{1.6}{4} \sum_{j=1}^{4} c_j.$$

Stage 3 Punishment

During the punishment stage subjects have the possibility to punish each other. At the beginning of the stage, all subjects[10] receive an endowment $e^P = 10$ Token to pay for

[8]The public good is called a project.

[9]Such a set-up constitutes a social dilemma, because subjects' strictly dominant strategy is to contribute 0 Token, however, the social surplus is maximized when everyone contributes the maximal possible amount of 20 Token.

[10]Subjects in treatments without punishment (NFnoP and noNFnoP) also received the endowment at the end of a period. This is also the case for the endowment received in stage 4.

exerted punishment and are informed of the contributions of the other group members. Subject i decides how many punishment points p_{ij} to assign to group member j. The income of the targeted group member j is reduced by $3 * p_{ij}$. For every assigned punishment point the punisher needs to bear costs of 1 Token. In other words, subjects have the possibility to reduce others' income by 3 Token by giving up 1 Token themselves. Subjects cannot spend more on punishment than their endowment e^P. [11] Subject i's payoff from stage 3 is given by

$$\pi_i^{III} = e^P - \sum_{j \neq i}^{4} p_{ij} - 3 * \sum_{j \neq i}^{4} p_{ji}.$$

Stage 4 Counter-Punishment

In this final stage subjects have the possibility to retaliate against their punishers. In order to fund their counter-punishment each subject receives an endowment of $e^{CP} = 5$ Token. Subjects are told by whom and by how much they were punished in the previous stage. If subject i received at least one punishment point ($p_{ji} > 0$) from group member j, then and only then i has the possibility to assign counter-punishment points cp_{ij} to j. Counter-punishment points have the same characteristics as punishment points, that is, they cost 1 Token to assign and they reduce the income of the recipient by 3 Token. Analogously, subjects cannot spend more on counter-punishment than their endowment e^{CP}. This design is especially hostile to high contributions, because counter-punishment can only be used to retaliate against punishers, in particular, low contributors can use this stage to retaliate against high contributors who try inducing high contributions. Generally, subjects have an incentive to strategically delay their punishment in order to avoid counter-punishment, since subjects can only assign counter-punishment if they received punishment from a certain group member, this is impossible in our setting. Furthermore, subjects generally have an incentive to punish others excessively in order to strip them of the funds to pay for counter-punishment. In our set-up, all subjects have the same counter-punishment power due to their endowment e^{CP}, which renders excessive punishment on stage 3 unnecessary. The payoff of subject i from stage 4 is given by

[11] This insures that all subjects have the same punishment possibilities, in particular, low contributors, who are richer than high contributors, do not have greater punishment power. Hence, there is no reason for subjects not to contribute in stage 2 in order to fund punishment on the next stage.

$$\pi_i^{IV} = e^{CP} - \sum_{j \neq i}^{4} cp_{ij} - 3 * \sum_{j \neq i}^{4} cp_{ji}.$$

At the very end of a period, subjects see once more an overview of the period. For every group member they see the contribution, the assigned and received punishment and counter-punishment points.

Total period profit is given by $\pi_i = \pi_i^{II} + \pi_i^{III} + \pi_i^{IV}$ and counts towards to the final payoff a subject. If a stage was not present in a given treatment, subjects would nevertheless receive the endowment from this stage to keep this aspect constant across treatments.

Methods and Procedures

We used a within-subject design, this means that in one session a subject participated in two treatments. Subjects were aware that the experiment consists of two parts and that they would remain in the same group, they however, first received only the instructions for their first treatment. After the initial treatment was finished the instructions for the second treatment were distributed. We conducted the experiment at two different computer laboratories. We started in May, June and October 2016 by running a total of 7 session in Zurich at the decision laboratory of the Department of Economics of the University of Zurich. Subjects, mainly student from the University of Zurich or the Swiss Federal Institute of Technology, were recruited using the software "hroot" (Bock et al., 2014). Recruited subjects had never participated in a public goods game before and did not study economics nor psychology. Subjects were invited for 90 minutes and to one session only. The experiment was conducted in German. We paid subjects according to the sum of all periods, negative period payoffs were possible, but the sum over all periods could not be negative, this rule never had to be applied. The exchange rate in Zurich was 10 Token = CHF 0.20, and subjects earned on average CHF 39.65, including a show-up fee of CHF 14.

We run an additional 8 sessions at the Centre for Decision Research and Experimental Economics (CeDEx) at the Universtity of Nottingham. Due to differences in the administrative process subjects in Nottingham were invited for 120 minutes instead of the 90 minutes in Zurich. Recruited subjects had not participated in an experiment that featured counter-punishment before. The experiment in Nottingham was conducted in English. The exchange rate in Nottingham was 10 Token = GBP 0.08, and subjects earned on average GBP 15.37, including a show-up fee of GBP 5.60. As we will outline in the

results section, the qualitative results hold for both subject pools, which underlines the robustness of our results. The experiment was programmed using z-tree[12] (Fischbacher, 2007). Table 3 gives an overview of all sessions including conducted treatments, location, number of subjects and date.

Table 3: Conducted Sessions

Treatment 1	Treatment 2	# subjects	Location	Date
noNF	NF	32	Zurich	17th May 2016, 13:15-14:45
noNF	NF	36	Zurich	7th June 2016, 15:30-17:00
noNF	NF	24	Nottingham	6th December 2016, 09:00-11:00
noNF	NF	28	Nottingham	6th December 2016, 14:00-16:00
NF	noNF	36	Zurich	17th May 2016, 15:30-17:00
NF	noNF	36	Zurich	7th June 2016, 13:00-14:30
NF	noNF	28	Nottingham	5th December 2016, 14:30-16:30
NF	noNF	24	Nottingham	6th December 2016, 11:30-13:30
noNF	noNF	36	Zurich	8th June 2016, 13:00-14:30
noNF	noNF	28	Nottingham	6th December 2016, 16:30-18:30
noNF	noNF	28	Nottingham	7th December 2016, 09:00-11:00
noNFnoP	NFnoP	36	Zurich	17th October 2016, 15:15-16:45
noNFnoP	NFnoP	28	Nottingham	7th December 2016, 16:30-18:30
NFnoP	noNFnoP	36	Zurich	17th October 2016, 13:15-14:45
NFnoP	noNFnoP	28	Nottingham	7th December 2016, 11:30-13:30

3 Analysis & Results

Our experiment allows analyzing the data in a between-subject manner, by examining only first treatments of sessions, but also in a within-subject way, by comparing the first treatment to the second treatment of a session. For better comprehensibility, we choose a combination of these two approaches by presenting the result based on pooled data for the main body of the paper. The appendix provides further analyses split in

[12]Due to a bug in the z-tree code, some subjects received a wrong information about received counter-punishment. This bug was only present in Zurich and in this sub-sample only 1.02% of all cases are affected.

between-subject and within-subject comparisons for both of our experimental locations, University of Zurich and University of Nottingham, separately. The conclusions drawn from these sub-samples are all qualitatively the same as those presented in this section. This procedure also provides us with with robustness checks of many of our results and conclusions.

The analysis is structured in two sections. We begin with an examination of the dynamics—formation and decay—of social norms and their compliance. The second part is concerned with the effect of social norms on behavior, that is, how they impact cooperation, punishment, counter-punishment and welfare.

Norm Formation, Norm Decay and Norm Compliance

Our fist concern is with the consensus regarding appropriate behavior, which we refer to as the strength of the norm. Recall, a social norm requires a *shared* understanding about appropriate actions. Whether or not punishment opportunities play a role in the emergence of a social consensus composes an empirical question. First, note that a difference in average demanded contributions does not imply a difference in consensus. It is, for instance, possible that without punishment the demanded contributions are lower, but as long as subjects agree to the same degree that these contributions levels are appropriate, there would be no difference in the consensus. So the question is how are punishment opportunities expected to impact the agreement among subjects. Without question, a perfect consensus can potentially emerge even when no punishment opportunities exist. A plausible candidate for a social norm with perfect consensus would be that everyone agrees that the maximal amount should be contributed. This social norm is as appealing with and without punishment. Normative request might not be oriented at this ideal case, but rather take into account what subjects actually do, that is, subjects may deem it appropriate to contribute less when group members contributed little previously and vice versa. In that case, one does also not expect an impact of punishment, since subjects always have the possibility to observe other group members' contributions.

Another plausible scenario is the following. Norm disobedience may stoke disagreement within a group. Some group members may uphold their demands of surplus maximizing contribution levels, whereas others may scale their normative requests down when they observe violations of the norm. If punishment is required for norm obedience, then we expect strong and stable norms under a sanctioning institution and decaying ones without. A final hypothesis centers around the effect of counter-punishment. Less demanding,

but potentially stronger, norms may arise due to the fact that counter-punishment is a constant threat when punishment is exerted. Demanding group members may lower their requests when they are subject to retaliation after punishing others for not meeting their requests. The analysis reveals that punishment opportunities indeed compose a decisive force in the formation of a social consensus about appropriate actions.

Result 1 (social consensus):

In the presence of a punishment opportunity, a stable and strong social consensus about the normatively appropriate contribution level manifests, while in the absence of such a punishment opportunity relatively large and stable disagreement emerges.

The vast majority, 72% of all subjects, in the punishment treatment make a normative request that is maximally 1 Token smaller or greater than the average in their group in the last period. The corresponding figure for groups without sanctioning possibilities is only 29%. These numbers clearly illustrate the widespread consensus formed in NF and prevailing disagreement in noNF. More formally, we take the coefficient of variation (CV) of subject's normative requests as a proxy for the degree of disagreement about appropriate behavior among subjects. This measure for the dispersion of a distribution is defined as the standard deviation divided by the mean.[13] Figure 9 depicts how the average CV develops over time. The graph illustrates that the degree of agreement about how much one ought to contribute is a lot more pronounced in NF compared to NFnoP despite roughly equal levels at the very start. The difference in groups' average CV over all 15 periods with norm formation opportunities is statistically significant (Wilcoxon rank-sum tests[14], n = 93 groups, p = 0.0003).

[13] We compute groups' CV = $\frac{\sqrt{\frac{1}{3}(\sum_{i=1}^{4}(Request-Norm)^2)}}{Norm}$

[14] We test the hypothesis $Pr(X_{treatment\ 1} > X_{treatment\ 2}) = Pr(X_{treatment\ 2} > X_{treatment\ 1})$, where $X_{treatment\ 1}$ is the outcome variable in one treatment and $X_{treatment\ 2}$ is the same variable in the other treatment. Treatment differences analyzed with the Wilcoxon rank sum test are all based on independent group averages of all 15 periods.

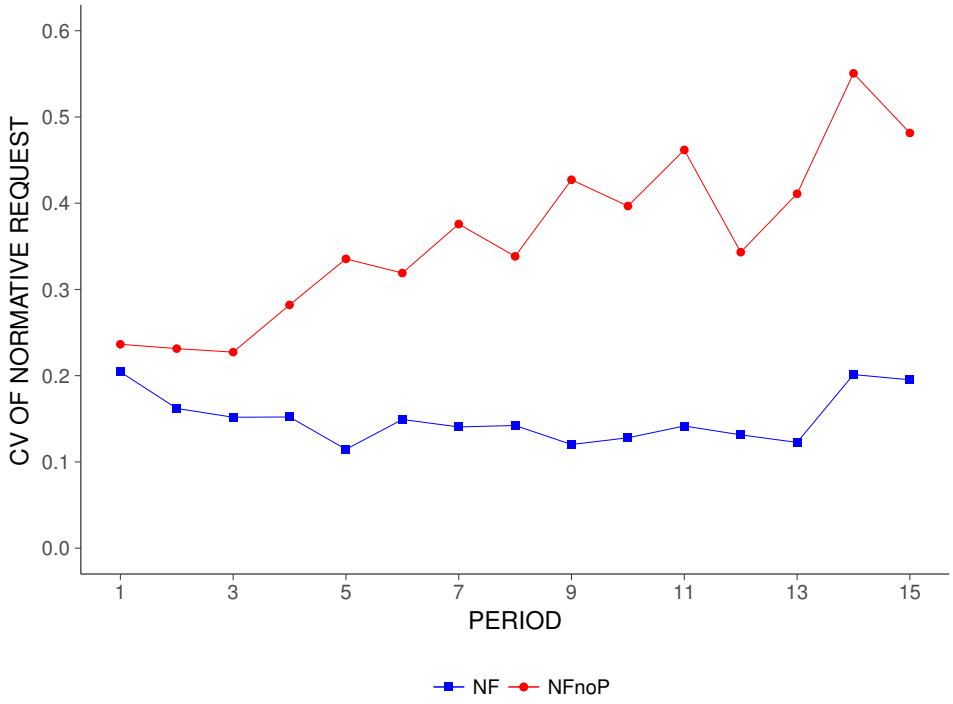


Figure 9: Coefficient of variation in normative requests over time (strength of norm)

Thus punishment induces indeed stronger social norms, however, as mentioned above this does not imply that these norms are also more demanding. Average requests could be lower despite being less dispersed. One reason stems from the fact that in our setting punishment opportunities are always accompanied by means of counter-punishment. Demanding subjects might initially punish those who contribute below their request, but these low contributors can always retaliate. This could signal the demanding subjects that no social norm was violated and that their punishment behavior was not in order. These subjects might lower their requests in turn. This way, punishment may lead to a higher consensus, but to less demanding social norms. On the other hand, without punishment, cooperation may not be sustained, and some subjects could lower their request in face of actual behavior. The pattern shown by the data suggests that counter-punishment does not have the power to make demanding subjects to adopt lower demands.

Result 2 (content of consensus):

In the presence of a punishment opportunity, the normative consensus quickly demands almost full cooperation while in its absence a more lenient average normative request emerges.

Figure 10 illustrates the content of the norm and its stability for the cases with and without punishment opportunities. The graphs show the evolution of the average content over time. In the first period, there is no significant difference in the content adopted by groups that have the possibility to punish compared to those that do not (Wilcoxon rank-sum test, n = 93 groups, p = 0.9735).[15] However, over the periods requested contributions in NF are slightly (+0.0021 per period) increasing, but, this increase is not significant (p = 0.1197).[16] On the contrary, without punishment, the content is significantly declining over time (-0.0079 per period; p = 0.0007).[17] This leads to significant differences when aggregating over all 15 periods with norm formation opportunities; requested contributions are significantly (Wilcoxon rank-sum test, n = 93 groups, p = 0.0005) higher when punishment is possible.[18] This shows that social norms are stable and demanding in the presence of sanctioning. In contrast, they are less stable and less demanding when punishment opportunities are absent.

[15]Insignificant first period differences between NF and NFnoP are also present in an OLS regression of contributions on the treatment dummy (p = 0.8061), with fixed effect for location and clustered standard errors on the group level.

[16]These results are based on an OLS regression of groups' norms on period, treatment dummy for NF and the interaction of these two with fixed effect for location and clustered standard errors on group level.

[17]see footnote 16

[18]Significant differences between NF and NFnoP are also present in OLS regressions of norm on treatment dummy (p = 0.0004), with fixed effect for location and clustered standard errors on group level.

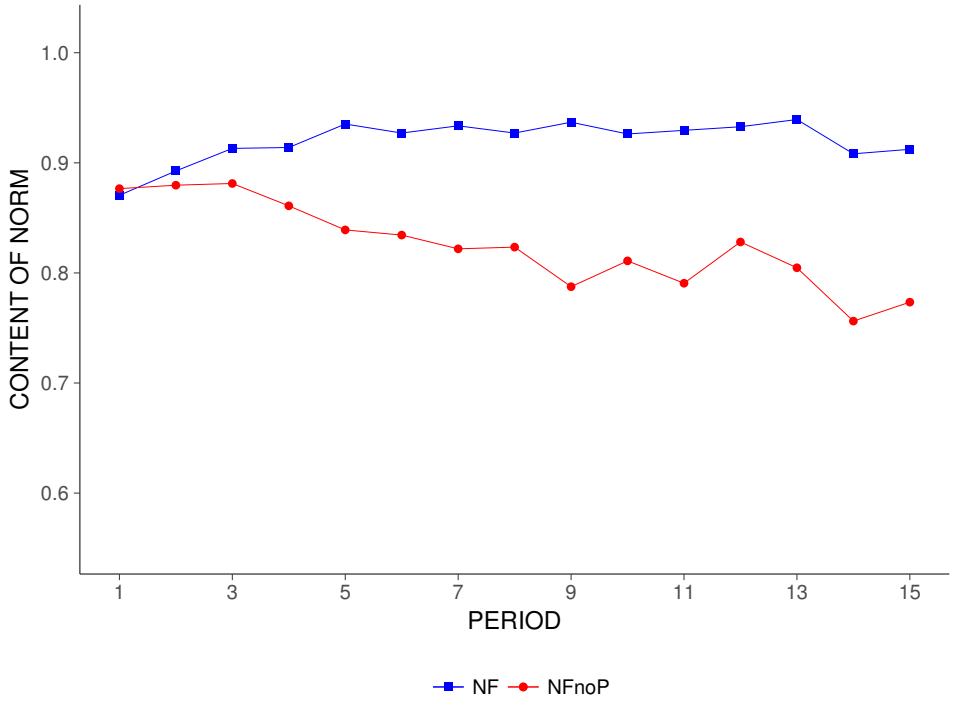


Figure 10: Average content of norm over time

Note: Norms are normalized to 1, i.e. 1.0 corresponds to a requested contribution of the maximal possible level of 20 Token.

Result 1 and 2 raise the question whether the norms that emerge in the two conditions really have behavioral traction or whether they are merely a form of cheap talk. In principle, subjects could reason that expressing what group members should do is one thing but actual behavior is another, because there is a substantial financial incentive to contribute nothing to the public good.

Result 3 (obedience to social norm):

(a) *When punishment is possible subjects, obey the demanding cooperation norm.*

(b) *In the absence of punishment opportunities, subjects regularly violate the relatively less demanding cooperation norm.*

Figure 11 shows subject's normalized average deviation from the norm, that is, actual minus requested contributions divided by requested contributions. This figure provides an illustration of how seriously subjects take the norm, and shows that whether or not the social norms is merely cheap talk depends on the possibility of subjects to punish each other. When there are punishment opportunities people live up to the prevailing social

norm although it demands very high cooperation levels. On the contrary, in the absence of punishment norm obedience strongly unravels over time although the norm is much less demanding. The difference is statistically highly significant (Wilcoxon rank-sum tests, n = 93 groups, p = 0.0000).[19] Thus, peer punishment is key for norms to be obeyed. Note also that the strong obedience to the social norm in the punishment condition holds despite the fact that punishers may fear counter-punishment, that is, norm obedience is obtained in an environment that is not very favorable for norm enforcement through punishment.

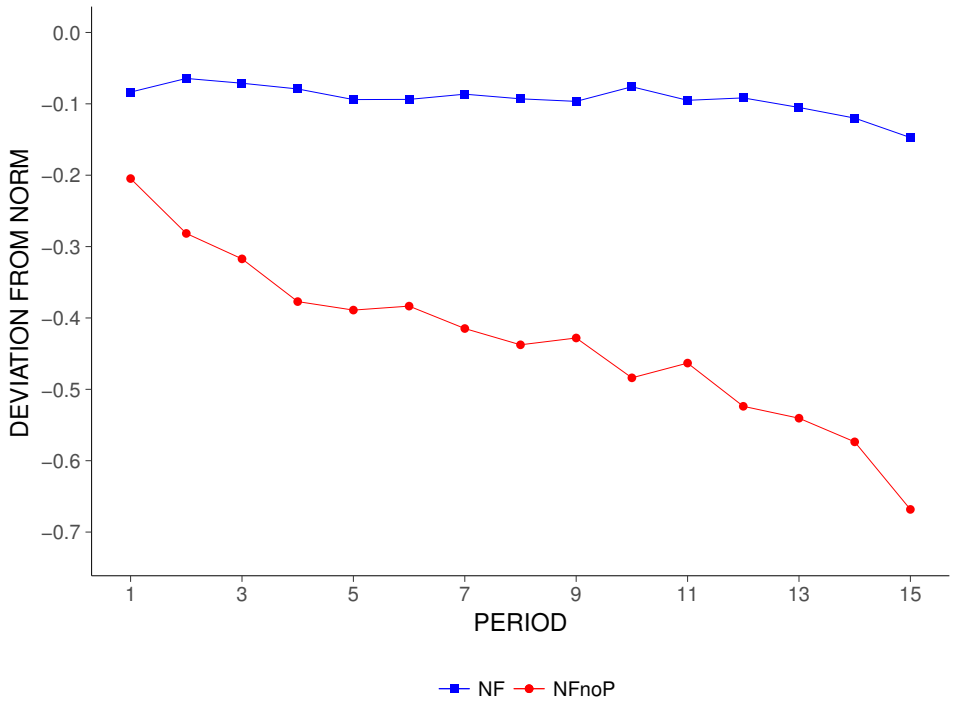


Figure 11: Deviations from norm over time

Note: Deviations from norm are normalized to 1. A deviation of 0 corresponds to a contribution that corresponds to the actually requested contribution, a -0.5 represents a contribution that is half of what was requested, and -1 indicates a contribution of 0 Token.

Discussion of the Dynamics of Norm Formation and Norm Decay

The dynamics of norm formation and norm decay warrants a closer examination. We therefore conducted a second set of experiments to answer the question whether the patterns described in the previous results indeed indicate changes in the social norm, and

[19] An OLS regression of normalized deviations from the norm on a treatment (NF=1 / NFnoP=0) dummy (+0.1633, p = 0.0037), the current period (-0.0254, p = 0.0000) and the interaction of these two predictors (+0.0220, p = 0.0000) underlines the results that in noNFnoP subjects violate the norm more strongly especially as time progresses. Standard errors are clustered on group level.

whether our norm formation device in fact allows the establishment of a social norm. Specifically, the new experiments investigate whether lower contributions to the public good become socially acceptable when group members request on average fewer contributions and the previous demands have been violated as it is the case in the NFnoP treatment.

We elicit the social norms using the Krupka-Weber method (Krupka & Weber, 2013). There are two treatment conditions: subjects either evaluate behavior in our NF (norm formation with punishment) or in our NFnoP (norm formation without punishment) treatment. Subjects read the original instructions of these experiments and answer the control questions. Subjects are then asked to evaluate the social appropriateness of certain contribution levels to the public in the first (t=1) and last (t=15) period. For the evaluation of the first period, subjects receive the information that according to the average opinion of the group each group member should contribute 0.9 (18 Token). This number corresponds to the average requested contribution in the first period in either treatment (see figure 10). This first evaluation is therefore the same for all subjects with the exception that half of them have read the instructions including punishment (NF) and the other half those without punishment (NFnoP). For the evaluation of the last period, we inform subjects of the actual behavior of one particular, qualitatively representative, group who participated in the public goods experiment. They receive information (graphical and written) about the average requested contributions over time and about average actual contributions over time. Subjects who evaluate the social norm in case there is no punishment, see that requested contributions decline from 0.9 to 0.4 (18 to 8 Token) and that actual contributions are between 0.25 and 0.5 (10 and 5 Token). Requested contributions in the last period, the period they need to evaluate, are 0.4 (8 Token). Subjects who read the instructions with punishment see that requested contributions are always between 0.9 and 1 (18 and 20 Token) and that actual contributions are between 0.85 and 1 (17 and 20 Token). They are also informed that requested contributions in the last period are 1 (20 Token).

For both periods, subjects evaluate three different contributions levels; high contributions of 0.9–1, medium contributions of 0.4–0.5, and low contributions of 0.15–0.25. For each contribution level, subjects select one of five possible social appropriateness ratings: "very socially inappropriate," "somewhat socially inappropriate," "neutral: neither socially inappropriate nor appropriate," "somewhat socially appropriate," and "very socially appropriate." Subjects therefore evaluate the following six different situations: high, medium, and low contributions each in the first and the last period. At the end of the experiment two of those situations are randomly chosen for payment. If a subject's

answer in a chosen situation corresponds to the most frequent answer given, this subject receives an additional payment of CHF 10 in addition to the show-up fee of CHF 15.

We conducted two experimental sessions in August 2018 with a total of 69 subjects (35 and 34 evaluating NF and NFnoP respectively). Subjects in either session were randomly assigned to evaluate NF or NFnoP. Sessions lasted 60 minutes and earnings were CHF 28.48 on average. The experiment was programmed in z-Tree (Fischbacher, 2007) and recruitment was implemented with hroot (Bock et al., 2014). The experiment was conducted in German at the decision laboratory of the Department of Economics of the University of Zurich.

Result 4 (dynamics of social norm):

(a) *Stable high normative requests and norm obedience solidify a social norm of highly cooperative behavior.*

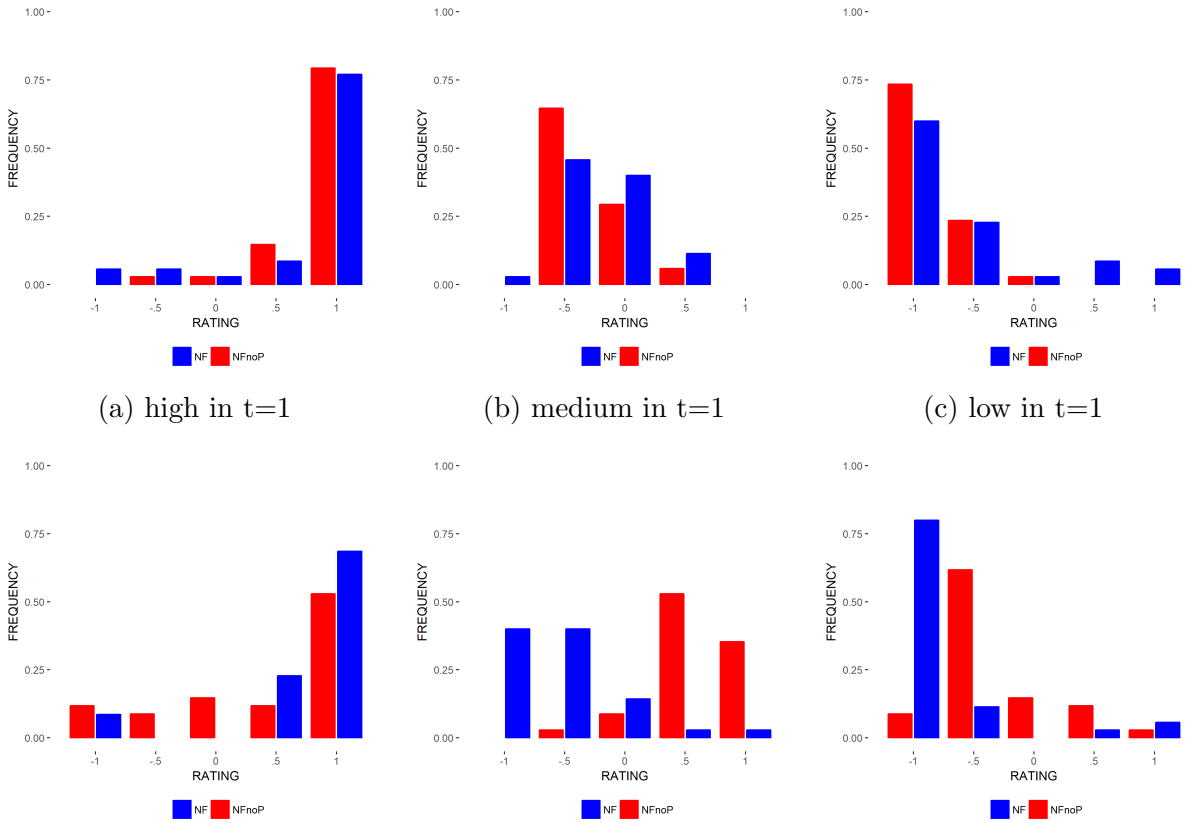
(b) *Declining normative requests and norm disobedience cause the decay of the social norm of cooperation.*

Figure 12 illustrates subjects' answers for each situation, the accompanying table lists the means of answers across treatments and the corresponding t-tests. The data reveals a clear social norm for the first period (figures (a)-(c)). High contributions at or close to the surplus maximizing level are highly socially appropriate, medium contributions are already somewhat inappropriate, and low contributions clearly violate appropriate social behavior. The existence of punishment opportunities makes little difference in this regard. In stark contrast, the social norm in period 15 strongly differs between the punishment and no punishment setting. The social norm in the NF setting solidifies over the course of the 15 periods. High contributions are still clearly socially acceptable, however, medium and low contributions become even less acceptable (p=0.000 and p=0.021 based on paired t-test). The social norm in the NFnoP setting on the other hand decays. High contributions are still acceptable, but markedly less so (p=0.005, paired t-test). Medium contributions become the social norm, that is, they are now even more socially acceptable than high contributions. This change in the appropriateness of medium contributions, that now correspond to the requested contributions, is statistically significant (p=0.000, paired t-test). Low contributions become considerably less socially inappropriate (p=0.000, paired t-test) in this scenario.

These changes in the social norms between the NF and the NFnoP setting mean that

in the last period there are now striking differences between the two settings. High contributions are somewhat more appropriate with punishment than without (p=0.072). Substantial changes occur for medium and low contributions, both of which are indisputably less appropriate with punishment than without (p=0.000 and p=0.000).

Figure 12: Dynamics of social norm across treatments



(a) high in t=1 (b) medium in t=1 (c) low in t=1

(d) high in t=15 (e) medium in t=15 (f) low in t=15

Treament Differences in Social Norm

Situation	Mean NF	Mean NFnoP	Difference	t-value	p-value
(a) high contribution in t=1	0.73	0.85	−0.12	−1.08	0.286
(b) medium contribution in t=1	−0.20	−0.29	0.09	1.16	0.252
(c) low contribution in t=1	−0.61	−0.85	−0.24**	2.11	0.039
(d) high contribution in t=15	0.71	0.43	0.28*	1.83	0.072
(e) medium contribution in t=15	−0.56	0.60	−1.16***	−11.26	0.000
(f) low contribution in t=15	−0.79	−0.31	−0.48***	−3.97	0.000

Notes: *p<0.1; **p<0.05; ***p<0.01; Figures (a)-(f) show the relative frequency of subjects' answers about their belief regarding the social appropriateness of contributing certain amounts to the public good. *High* refers to contributions of 0.9–1.0 (18–20 Token), *medium* refers to contributions of 0.4–0.5 (8–10 Token), and *low* refers to contributions of 0.15–0.25 (3–5 Token). Social appropriateness ratings are converted to numbers; "very socially inappropriate" = −1, "somewhat socially inappropriate" = −0.5, "neutral: neither socially inappropriate nor appropriate" = 0, "somewhat socially appropriate" = 0.5, "very socially appropriate" = 1. The table depicts, for all six situations and for both treatments, the average answer, the difference between the two treatments, the t-value for the hypothesis that the difference is zero, and the corresponding p-value.

Result 3 does not yet fully rule out that the social norms summarized in results 1, 2, and 4 are merely cheap talk, because it could be possible that subjects achieve the same cooperation levels even in the absence of these social norms. For this reason, we compare the cooperation and punishment patterns in the treatments with a norm formation opportunity to those without such a device in the following results. We want to start with a discussion of the potential reasons for why the norm formation opportunity may affect cooperation rates. One reason could be that subjects in fact form a high cooperation norm and some subjects may be intrinsically motivated to comply with this norm, which would lead to higher cooperation rates relative to a situation with no formation opportunity. Intrinsic motivation might originate from feeling guilt for not complying with the prescription (Elster, 1989a).

A second channel could be due to the well-documented (e.g. Fischbacher et al., 2001) existence of "conditional cooperators", that is, of subjects who are willing to make high contributions if they expect other members to make high contributions. Thus, if the norm formation opportunity facilitates a high social cooperation norm conditional, cooperators may expect higher contributions from other group members. This would in turn induce them to make higher contributions. Third, Bicchieri (2006) reasons that one mechanism of social norm compliance is "that one accepts others' normative expectations as well founded." Non-compliance would in that case require the subject to justify his deviation, even if only to himself, by offering better reasons than those of the other group members. Fourth, forming a social norm may reduce the appropriateness of relatively low contributions and this may enhance the legitimacy and, hence, the credibility of punishing free-riders. If this channel is operative it might even be the case that we observe higher cooperation levels when norm formation is possible, although one does not observe higher punishment levels.

We suggest a final potential channel, which is derived from the well established result that punished free-riders subsequently increase their contributions. In the presence of a norm formation device, this increase could be more pronounced, because free-riders regard their received punishment as more legitimate. Note that the first three potential channels for higher cooperation levels under a norm formation device can become operative regardless of whether subjects have a punishment opportunity. Intrinsic motivations to obey a high cooperation norm and the presence of conditional cooperators could increase cooperation levels also when punishment is not possible. However, in clear contrast to these predictions, we observe the following:

Result 5 (cooperation):

(a) *When there are punishment opportunities, the introduction of a norm formation device unambiguously increases cooperation. This increase occurs immediately, that is, in period 1 and is maintained throughout the experiment.*

(b) *In sharp contrast when punishment is not possible the norm formation device is completely ineffective.*

Result 5(a) is illustrated in Figure 13, which shows the impact of the norm formation opportunity on cooperation rates. The figure indicates that the norm formation opportunity causes sizable increases in average cooperation rates and that the difference between NF and noNF remains fairly stable over time. In addition, the figure also shows that the cooperation enhancing role of the formation opportunity becomes effective immediately after its introduction: if one compares the first round of noNF with the first round of NF, one observes an increase in cooperation rates of roughly 15 percentage points. Both differences are statistically significant.[20]

[20]We test the following way. An OLS regression with a treatment dummy for NF and a fixed effect for location with standard errors clustered at the group level. The p-value for the treatment dummy is p=0.0026 over all periods and p=0.0009 for the first period alone. For this analysis we exclude the data of the second treatment when the first treatment was NF, due to spillover effects—a norm was formed—from this treatment to the second one. A non-parametric test is not suitable in this situation due to our data structure that features dependent and independent observations at the aggregate group level. The appendix provides separate results for within-subjects and between-subjects comparisons and for each location based on both parametric and non-parametric tests. All conclusions are qualitatively the same as the one shown here for the pooled data.

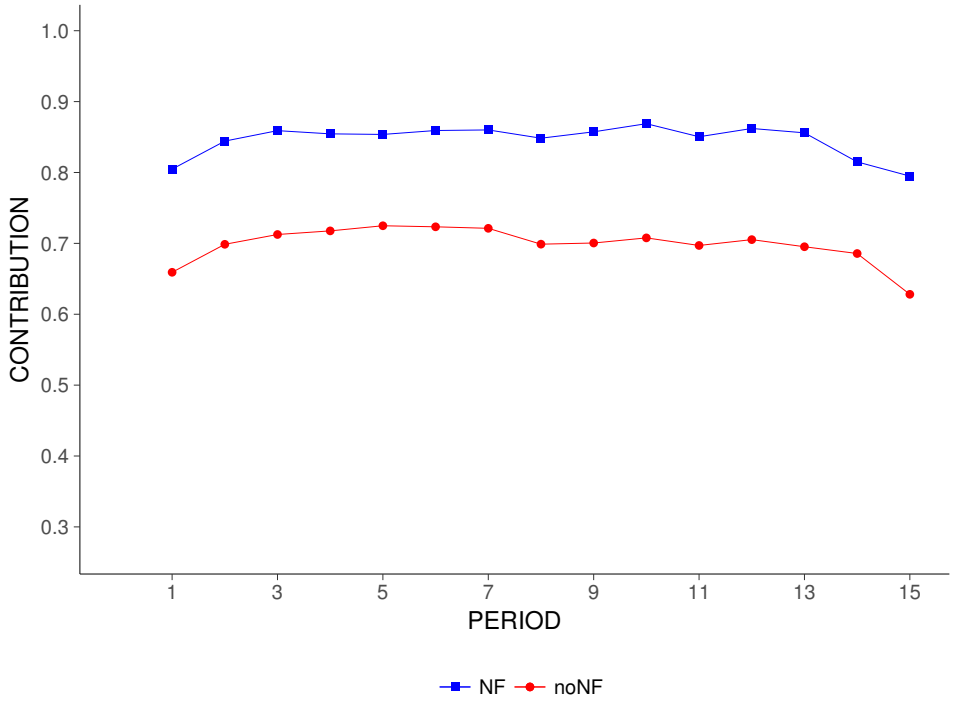


Figure 13: Contributions with punishment

Note: Contributions are normalized to 1, i.e. 1.0 corresponds to a contribution of the maximal possible level of 20 Token.

This pattern contrast sharply with the pattern of cooperation in the absence of a punishment opportunity (Result 5(b)). Figure 14 shows that cooperation rates quickly and strongly unravel with the norm formation opportunity without punishment (NFnoP). In addition, the cooperation levels in NFnoP are—with the exception of the first period—even *lower* than those in noNFnoP. Average contributions with the norm formation opportunity are about 6 p.p. lower (not significant) when punishment is not possible, despite the initially significantly 11 p.p. greater contributions in this treatment.[21]

[21] We test the following way. An OLS regression with a treatment dummy for NFnoP and a fixed effect for location with standard errors clustered at the group level. The p-value for the treatment dummy is p=0.2307 over all periods and p=0.0362 for the first period alone. For this analysis we exclude the data of the second treatment when the first treatment was NFnoP, due to spillover effects—a norm was formed—from this treatment to the second one. A non-parametric test is not suitable in this situation due to our data structure that features dependent and independent observations at the aggregate group level. The appendix provides separate results for within-subjects and between-subjects comparisons and for each location based on both parametric and non-parametric tests. All conclusions are qualitatively the same as the one shown here for the pooled data.

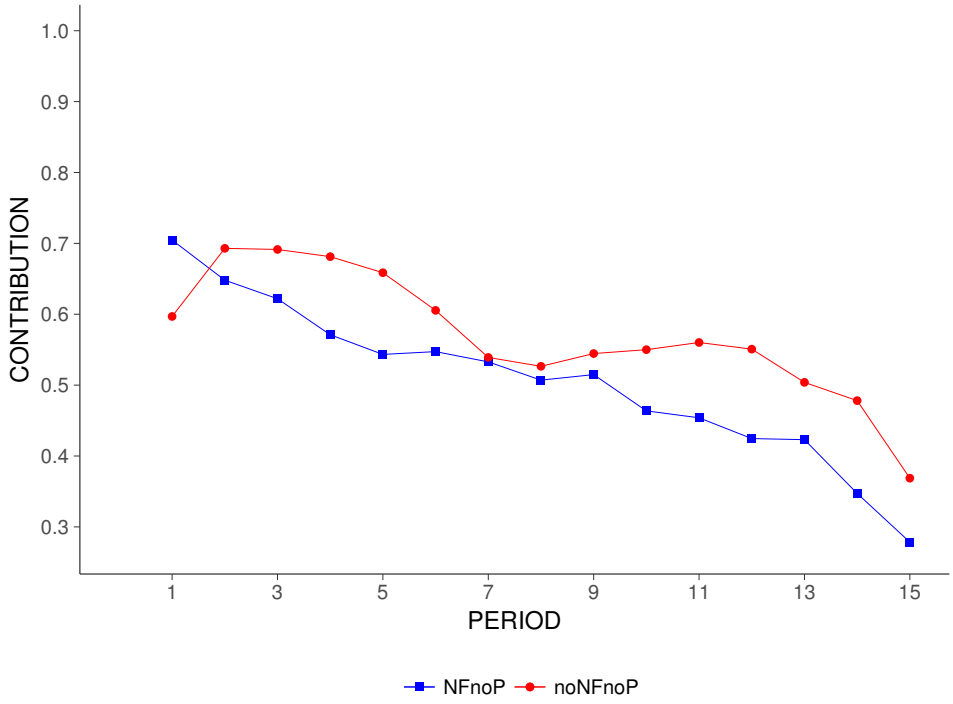


Figure 14: Contributions without punishment

Note: Contributions are normalized to 1, i.e. 1.0 corresponds to a contribution of the maximal possible level of 20 Token.

Next, we want to shed some further light on the causal relationship between social norms and contributions by disentangling direct and indirect (via reaction to punishment) effects by means of regression analysis. Table 4 depicts six different OLS estimations of contribution. All variables are described in detail in table 7 in the appendix. We estimate separate regressions for each treatment. Regressions 'NF (1)' and 'NFnoP (1)' show that the norm strongly predicts the actual contribution made by the subject without controlling for anything expect a location FE and the number of periods. This relationship is stable when norms can be enforced by means of punishment, but diminishes when no such possibilities exist (see coefficient for *Norm * Period*, this further provides evidence for Result 3). Regressions 'NF (2)' and 'NFnoP (2)' show that the norm has predictive power even when controlling for established predictors of contributions. In these regressions we additionally control for other group members' average contribution in the previous period (conditional cooperation) and the punishment received by the subject one period before (reaction to punishment). Meaningfully estimating the latter requires that subject's previous contribution is included in the regression since free-riders naturally attract more punishment. The aforementioned trends in time for the traction of the norm is confirmed in these richer models; without punishment opportunities the norm loses its behavioral

traction over the course of the 15 periods. With punishment, an increase in the requested contributions of 1 Token leads on average[22] to an increase of 0.300 in actual contributions c.p.. This is an increase of similar magnitude as our estimation for conditional cooperation (+0.306), that is, the predicted increase of a subject's contribution if all other group members had contributed 1 more Token in the previous period.

We take this as evidence of the direct effect social norms have on behavior when enforcement is possible. Regressions 'NF (3)' and 'noNF (3)' are included to test our fourth proposed channel, namely that subjects' reaction to punishment might differ across treatment. The increase in contributions due to received punishment is captured by the coefficient for *Punishment [t-1]* in the regressions shown in table 4. Comparing the coefficients in 'NF (3)' and 'noNF (3)' shows that subjects increase their contributions less after received punishment when there is no norm formation opportunity. This difference in coefficients is marginally significant (p = 0.0682).

[22] Average effect when period in treatment t=8.

Table 4: Regression Contribution

	Dependent variable: Contribution					
	NF (1)	NFnoP (1)	NF (2)	NFnoP (2)	NF (3)	noNF (3)
Norm	1.194***	1.627***	0.348***	0.468***		
	(0.140)	(0.165)	(0.067)	(0.104)		
Norm * Period	0.005	−0.041***	−0.006	−0.025***		
	(0.011)	(0.015)	(0.005)	(0.008)		
Contribution [t-1]			0.553***	0.380***	0.583***	0.490***
			(0.055)	(0.043)	(0.052)	(0.023)
Others' average contribution [t-1]			0.306***	0.439***	0.395***	0.495***
			(0.051)	(0.052)	(0.052)	(0.023)
Punishment [t-1]			0.420***		0.424***	0.275***
			(0.078)		(0.077)	(0.054)
Period	−0.162	0.395	0.051	0.348***	−0.054***	−0.049***
	(0.218)	(0.243)	(0.104)	(0.119)	(0.014)	(0.012)
Constant	−5.606**	−14.313***	−3.819***	−6.122***	0.511*	0.335**
	(2.668)	(2.939)	(1.150)	(1.420)	(0.297)	(0.146)
FE location and treatment order	YES	YES	YES	YES	YES	YES
Observations	3,660	1,920	3,416	1,792	3,416	5,992
R^2	0.470	0.376	0.749	0.622	0.738	0.794

Note: *p<0.1; **p<0.05; ***p<0.01; OLS regression of subjects' contributions (0-20 Token) with clustered standard errors on group level. The first row of each column shows the models name, e.g. 'NF (2)', which shows on what treatment the data is based and the regression specification in brackets. The variables have the following meaning (range of possible values in brackets). Norm (0-20) is the number of Token requested as seen by subject in current period. Period (1-15) is the current period in this part of the experiment. * indicates interaction of variables. Contribution [t-1] (0-20) is this subject's contribution in the previous period. Punishment [t-1] (0-30) is this subject's total number of received punishment points in previous period. FE location and treatment order means that a fixed effect for the location University of Zurich and a fixed effect for first treatment in session was included.

Result 5 indicates that social norms cause higher contributions when peer punishment is possible. However, more contributions do not necessarily mean that the welfare of the group is higher because they may be associated with high levels of costly punishment

and counter-punishment. Next, we examine whether norm formation opportunities also increase the average welfare of the groups. Furthermore, there is an ongoing debate whether or not a peer-to-peer punishment scheme actually improves group welfare (e.g. Herrmann et al., 2008; Gächter et al., 2008). Despite the higher contributions to the public good, it is often the case that welfare is similar or even lower under such a punishment setting, because sanctions are costly for the punisher and the punished. However, these settings generally lack an opportunity to form a social norm, which we deem a crucial and omnipresent feature of many important settings of everyday life.

Result 6 (group welfare):

(a) *When punishment is possible the norm formation opportunity causes on average a significant increase in the realized gains from cooperation. Moreover, this increase occurs immediately, that is, it is present already in period 1.*

(b) *When punishment is not possible the norm formation opportunity does not affect the realized gains from cooperation.*

(c) *A decentralized peer-to-peer punishment scheme improves group welfare, but only when norm formation opportunities exists, despite the prevailing threat of counter-punishment. Without norm formation, there is no positive effect of punishment opportunities on welfare.*

In order to analyze group welfare we study the fraction of realized potential gains from contributions.[23] The potential gains from contributions of one subject are given by $e^C * (4*0.4-1) = 12$ Token. For these to be realized, subjects first need to actually contribute to the public good, and second the benefits from the public good must not be vanquished by the total costs of punishment and counter-punishment.

Figure 15 shows how the fractions of realized potential gains evolve over time with and without the norm formation opportunity when punishment possibilities exist. Figure 16 shows the same trajectories when no such punishment opportunities are in place. Note that in NFnoP and noNFnoP realized potential gains are just a transformation of contributions, hence, the results about treatments differences are the same as for contributions, that is, no increase in efficiency without punishment. Regarding treatments with punishment, the realized potential gains are roughly 23 p.p. (2.76 Token per group member and period) greater when the device to form a norm was present. This increase

[23]Note that the variable we look at is merely a positive linear transformation of subjects' earnings.

in groups welfare due to the norm formation opportunity takes place right from the beginning (+25 p.p. in period 1). These differences are statistically significant.[24] There are no welfare benefits of a decentralized peer-to-peer institution when subjects cannot form a norm at the beginning of a period. On average, subjects' realized potential gains from cooperation are 11 p.p. less (not significant) in noNF compared to noNFnoP, that is, when they have the opportunity to sanction one another, but no means to form a norm. With a norm formation opportunity on the other hand, they earn more under peer punishment. Realized potential gains are on average significantly 18 p.p. greater in NF compared to NFnoP.[25] The norm formation opportunity renders peer punishment unambiguously superior in terms of social welfare.

[24] We test the following way. An OLS regression with a treatment dummy for NF and a fixed effect for location with standard errors clustered at the group level. The p-value for the treatment dummy is p=0.0027 over all periods and p=0.0624 for the first period alone. For this analysis we exclude the data of the second treatment when the first treatment was NF, due to spillover effects—a norm was formed—from this treatment to the second one. A non-parametric test is not suitable in this situation due to our data structure that features dependent and independent observations at the aggregate group level. The appendix provides separate results for within-subjects and between-subjects comparisons and for each location based on both parametric and non-parametric tests. All conclusions are qualitatively the same as the one shown here for the pooled data.

[25] We test the following way. OLS regressions with a treatment dummies for either NF or noNF and a fixed effect for location with standard errors clustered at the group level. The p-value for the treatment dummy is p=0.0064 when comparing NF with NFnoP and p=0.2761 for the comparison of noNF with noNFnoP. For these analyses we exclude the data of the second treatment when the first treatment was NFor NFnoP, due to spillover effects—a norm was formed—from this treatment to the second one. A non-parametric test is not suitable in this situation due to our data structure that features dependent and independent observations at the aggregate group level.

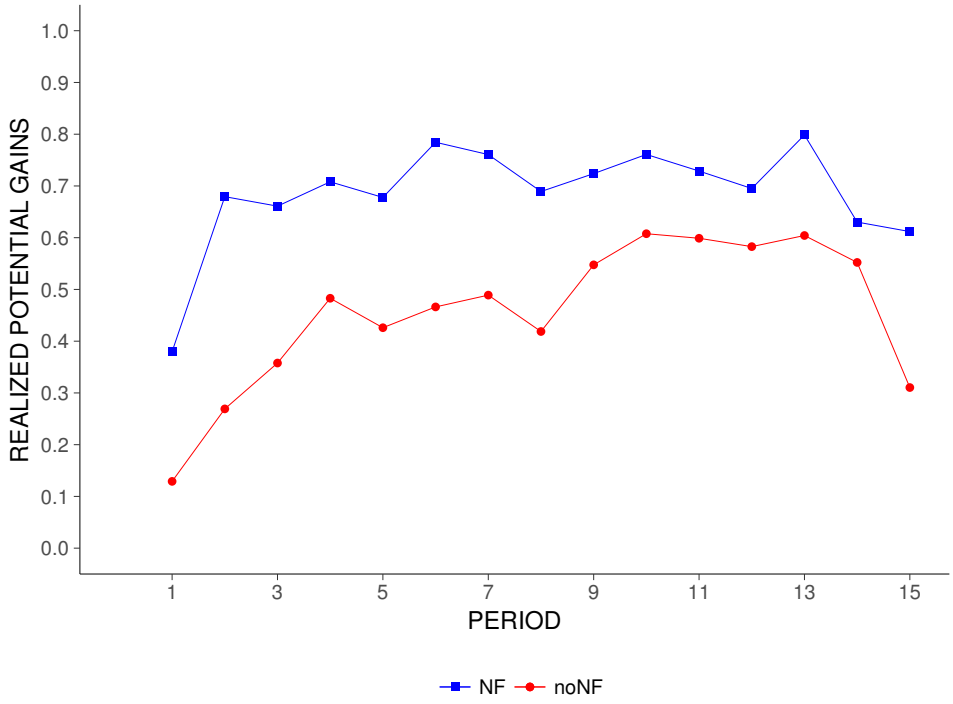


Figure 15: Group welfare with punishment

Note: Realized Potential Gains are normalized to 1, i.e. 1.0 corresponds to a realized gain of the maximal possible level of 12 Token.

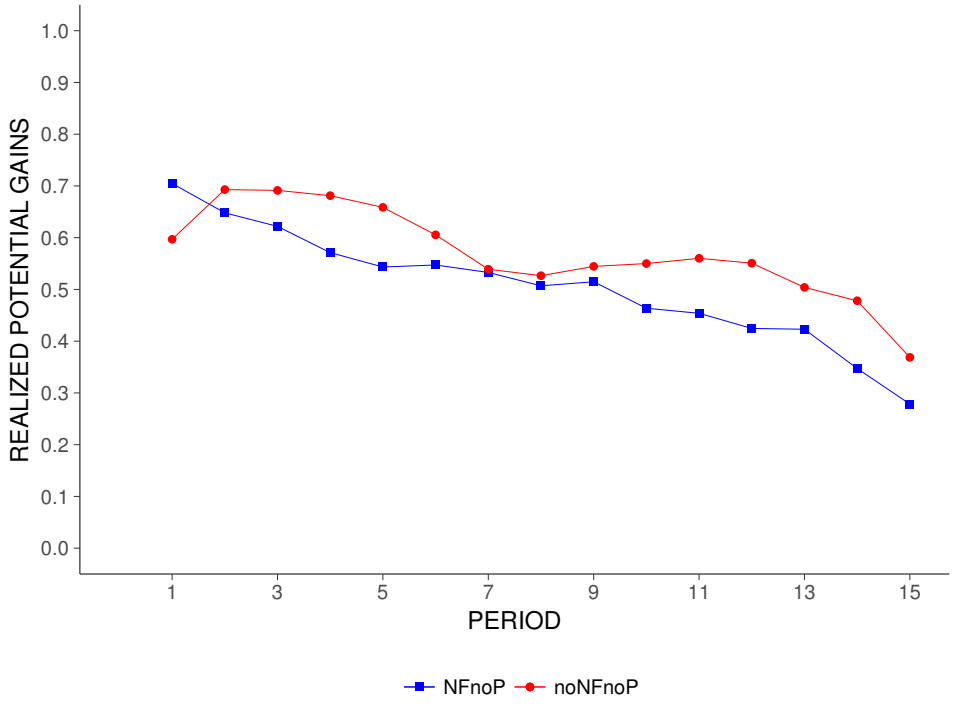


Figure 16: Group welfare without punishment

Note: Realized Potential Gains are normalized to 1, i.e. 1.0 corresponds to a realized gain of the maximal possible level of 12 Token.

We want to address the question of how the possibility to form a norm affects punishment behavior, now that we have established that the impact of the explicit norm formation device on contributions and group welfare crucially depends on the presence of norm enforcement opportunities, and that the efficiency gains from punishment hinge on an opportunity to form a norm. Arguably social norms legitimize the punishment of free-riders more strongly. Following this, we hypothesize that two opposing effects are plausible that drive whether free-riders receive more or less punishment under norm formation. First, more legitimate punishment could mean that less of it is required to sustain high cooperation rates. Second, and contrary, the higher legitimacy might induce subjects to punish more severely. Another common observed phenomenon is the prevalence of anti-social punishment, that is, punishment, which is directed towards group members that contributed relatively many Token to the public good. Social norm potentially undermines the legitimacy of the punishment of those who contributed more than the average, and could therefore remedy this problem. The data reveals the following:

Result 7 (punishment):

The norm formation opportunity leads to changes in punishment behavior by

(a) *significantly reducing the severity of the punishment of free-riders without lowering cooperation rates,*

(b) *slightly reducing the antisocial punishment of contributions above others' average contribution,*

(c) *increasing the overall punishment severity with higher average contributions of the potential punishers.*

Figure 17 depicts average received punishment points conditional on the deviation from average contribution of others. The graph shows that received punishment is greater in noNF compared to NF for every bin, except the one capturing subjects who contributed close to the average. In addition the graph pictures an increasing severity of punishment the stronger the deviation in either direction when there is no explicit norm formation opportunity. In NF this is only the case for negative deviations, this means, the tendency to increase punishment of those who contributed above average seems to be absent in NF, which stands in contrast to noNF. We rely on regression analysis in order to control for several factors that might confound the conclusions drawn form figure 17. Table 5 shows three OLS regression models for *Punishment probability* (i.e. whether or not punishment of exerted) and for *Punishment severity* (i.e. the amount of assigned punishment points if punishment was exerted) each.

There is reason to run the analysis for these two outcomes separately. First, intuitively the process to determine whether to punish at all potentially differs from the one about the extent of punishment given a positive decision has been reached. Secondly, more technical in nature, in the vast majority (96.2%) of cases no punishment is exerted. We include the average contribution to check whether punishment is correlated with the extent of cooperative behavior, that is, does punishment increase or decrease with the level of contributions. Previous research has identified the target's deviation in contributions from the rest of the group as an important predictor of punishment. We distinguish between positive and negative deviations. We also include the received counter-punishment in the previous period in order to see how counter-punishment affects punishment behavior . This calls for controlling assigned punishment in the previous period due to the fact that those who are willing to assign punishment are also the potential target of retaliation. Furthermore, we include controls for the period and a fixed effect for the location. Table

8 lists all predictors and their definition.

The regression confirms the conclusions drawn from the graph about the difference in punishment severity for free-riders. The coefficient for *Target's neg. deviation from others' average contribution* is negative for both treatments, but in noNF the coefficient is significantly ($p = 0.0181$) larger. Hence, groups in NF mange to sustain *higher* cooperation rates by punishing free-riders *less* severely. One explanation for this is that free-riders increased their contributions more for a certain amount of received punishment (as seen in table 4). Further, the regression reveals that in noNF, received punishment is significantly increasing in positive deviations. This is evidence for the prevalence of anti-social punishment. The coefficient for *Target's pos. deviation from others' average contribution* is positive and significant without norm formation (regression 'noNF (S)'). This increase of received punishment in positive deviations seems absent in 'NF (S)'. When groups can form a norm first, this coefficient is much smaller and insignificant. Hence, social norms potentially mitigate anti-social punishment. Notwithstanding, the difference in the two coefficients only has a p-value of $p = 0.1408$. Finally, the regression also provides evidence for result 7(c). The coefficient for *Average contribution w/o target* is significant and positive, that is, the more the rest of the group contributed on average the more severe the punishment under norm coordination ('NF (S)'). Without norm formation ('noNF (S)'), the relationship goes in the opposite direction, however, the coefficient is statistically insignificant. The difference between the coefficients is significant at the 1%-level ($p = 0.0065$) though. We argue that a positive relationship means that groups enforce further increases in contribution more and more the higher contributions get, since received punishment encourages higher contributions (see table 4). With norm formation, groups, therefore, seem to push for the really high contributions by even increasing the punishment severity the higher average contributions get.

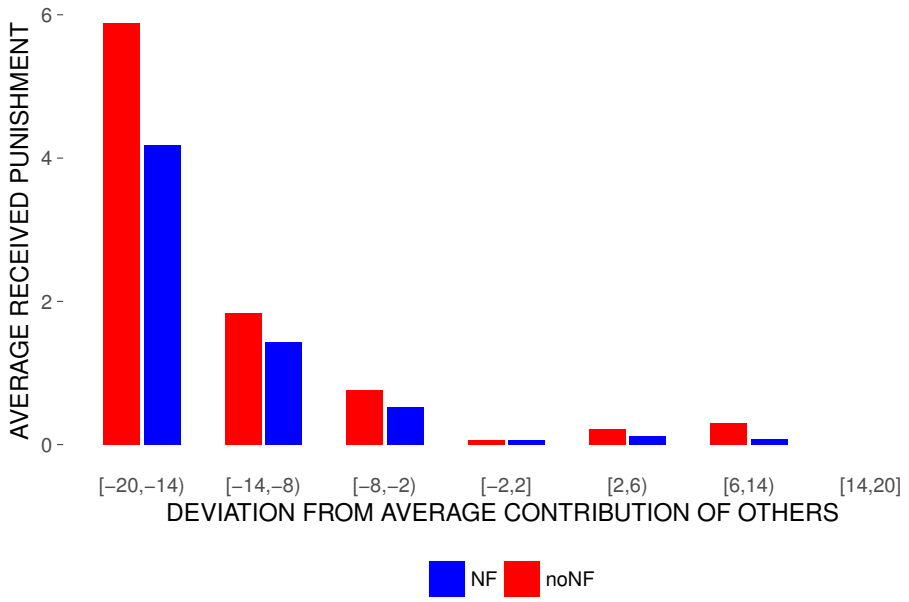


Figure 17: Average received punishment for levels of deviation from group

Table 5: Regression Punishment

	Dependent variable:			
	Punishment probability		Punishment severity	
	noNF (P)	NF (P)	noNF (S)	NF (S)
Average contribution w/o target	0.0001	0.001	−0.041	0.104**
	(0.0004)	(0.001)	(0.029)	(0.049)
Target's pos. deviation from others' average contribution	0.002	0.002*	0.183**	0.065
	(0.001)	(0.001)	(0.074)	(0.050)
Target's neg. deviation from others' average contribution	0.029***	0.030***	0.225***	0.138***
	(0.003)	(0.005)	(0.021)	(0.029)
Rec. counter-punishment [t-1]	0.006	0.001	−0.020	−0.153
	(0.004)	(0.005)	(0.067)	(0.118)
Assigned punishment [t-1]	0.026***	0.016***	0.214***	0.388***
	(0.005)	(0.004)	(0.063)	(0.078)
Period	−0.0004	−0.001**	0.034	0.125***
	(0.0004)	(0.001)	(0.029)	(0.034)
Constant	−0.005	−0.009	1.078**	−1.117
	(0.008)	(0.017)	(0.457)	(0.972)
FE location and tratment order	YES	YES	YES	YES
Observations	17,976	10,248	736	334
R^2	0.222	0.190	0.275	0.285

Note: *p<0.1; **p<0.05; ***p<0.01; OLS regression of punishment (probability models: 0 if 0 points assigned, 1 else; severity models: 1-10 punishment points) with clustered standard errors on group level; Based on data from periods $2-15$ of treatments NF and noNF. The models for punishment probability are based on all observations, those for punishment severity are based on the sub-sample of observations with a positive punishment decision. The first row of each column shows the models name, e.g. 'NF (P)', which shows on what treatment the data is based and the dependent variable (P) stands for probability model, (S) for severity model. The variables have the following meaning (range of possible values in brackets). Average contribution w/o target (0-20 Token) is the number of average contributed Token in the group excluding the subject that the punishment decision is aimed at (so called target). Target's pos. deviation from others' average contribution (0-20 Token) indicates how many Token the target has contributed above the average of the other three group members, 0 if contributed less. Target's neg. deviation from others' average contribution (0-20 Token, always non-negative) defined as how many Token the target has contributed below the average of other three group members, 0 if contributed more. Rec. counter-punishment [t-1] (0-15 counter-punishment points) is the number of received counter-punishment points of this subject in the previous period. Assigned punishment [t-1] (0-10 punishment points) number of punishment points this subject has assigned in total in the previous period. Period (1-15) is the current period in this part of the experiment. FE location and treatment order means that a fixed effect for the location University of Zurich and a fixed effect for first treatment in session was included.

There are 1279 cases of punishment and hence the same number of cases that could potentially trigger counter-punishment. We find that subjects made ample use of retaliation and stroke back at their punisher in 630 cases, which is just a little under 50% of all opportunities. What are the determinants of counter-punishment? Counter-punishment may be used for two distinct ends, reciprocity and strategic concerns. Reciprocating means that the probability and severity of counter-punishment is increasing in how unfairly a subject feels treated by the received punishment. Hence, more received punishment should result in more likely and more severe counter-punishment. Additionally, we would expect subjects to exert less and milder counter-punishment the fewer they contributed compared to the rest of the group, and the retaliate more often and more strongly the more they contributed above others' contribution. This is based on subjects taking into account why they were punished. Receiving punishment because of free-riding is then no longer considered that unfair or unkind, but seen even more so if received for giving more than the average group member. On the other hand, the strategic use of counter-punishment serves the purpose to deter future sanctions. If this drives counter-punishment, then strong free-riders should make more use of counter-punishment, because they potentially profit to a greater extent from defending their low contribution level.

What role do social norms play in this? Strategic concerns are not expected to be affected by the norm formation opportunity, however, reciprocity might. The norm formation opportunity could help subjects to understand why they received punishment. Being punished as a free-rider (cooperator) would then seem less (more) unkind, and subsequently lead to less (more) counter-punishment. To test these hypothesis we once more run a regression analysis. We regress *Counter-punishment probability* (i.e. whether or not counter-punishment was exerted) and *Counter-punishment severity* (i.e. how severe counter-punishment was given that a positive decision was reached) on several predictors. Table 9 lists all predictors and their meaning used in the regression. The regressions reveal the following patterns:

Result 8 (counter-punishment):

(a) *Counter-punishment becomes smaller the more the counter-punisher deviated negatively from others' average contribution—an effect that is slightly more pronounced in the norm formation treatment.*

(b) *Above average contributors are significantly more likely to counter-punish under a norm formation device.*

Coefficients for *Counter-punisher's neg. deviation from others' average contribution* are significantly negative both in 'noNF (S)' and 'NF (S)'. The coefficient under norm formation opportunities is more pronounced, the difference is, however, not statistically significant ($p = 0.1760$). So subjects retaliate less severely the more they deviate negatively from the average of the group. This indicates that they usually understand why they are punished and in consequence accept the received punishment, especially when norm formation is enabled. Regarding the counter-punishment decision (result 7(b)), the coefficient for *Counter-punisher's pos. deviation from others' average contribution* is negative in 'noNF (P)', but positive in 'NF (P)'. Neither is significantly different from zero, their difference, however, is statistically significant at the 5%-level ($p = 0.0270$), which provides evidence that above average contributors in NF are less inclined to accept their arguably uncalled punishment and regard it as more unkind compared to noNF, and therefore retaliate more often. The coefficient for *Received punishment* is also positive in all regressions, which provides further evidence that reciprocity, and not strategic concerns, play the decisive role in counter-punishment. One reason for this might be that the strategical use of counter-punishment seems to be ineffective. Consider again the punishment regression in table 5. The coefficient for *Rec. counter-punishment [t-1]* is never significantly different from zero indicating that counter-punishment does not deter punishment.

Table 6: Regression Counter-Punishment

	Dependent variable:			
	Counter-punishment probability		Counter-punishment severity	
	noNF (P)	NF (P)	noNF (S)	NF (S)
Counter-punisher's pos. deviation from others' average contribution	−0.018	0.020	0.110***	0.054
	(0.011)	(0.015)	(0.038)	(0.049)
Counter-punisher's neg. deviation from others' average contribution	−0.012***	−0.008*	−0.038***	−0.071***
	(0.005)	(0.005)	(0.014)	(0.021)
Received punishment	0.031***	0.016	0.307***	0.300***
	(0.011)	(0.011)	(0.022)	(0.039)
Period	−0.011**	−0.014**	0.010	0.031
	(0.005)	(0.007)	(0.014)	(0.025)
Constant	0.530***	0.557***	1.438***	1.602***
	(0.062)	(0.094)	(0.195)	(0.376)
FE location and tratment order	YES	YES	YES	YES
Observations	869	410	412	218
R^2	0.039	0.047	0.385	0.296

Note: *p<0.1; **p<0.05; ***p<0.01; OLS regression of counter-punishment (probability models: 0 if 0 points assigned, 1 else; severity models: 1-5 counter-punishment points) with clustered standard errors on group level; Based on data from periods $1-15$ of treatments NF and noNF. The models for counter-punishment probability are based on all observations where the subject received at least one punishment point from the target of the counter-punishment, those for punishment severity are based on the sub-sample of observations with a positive counter-punishment decision. The first row of each column shows the models name, e.g. 'NF (P)', which shows on what treatment the data is based and the dependent variable (P) stands for probability model, (S) for severity model. The variables have the following meaning (range of possible values in brackets). Counter-Punisher's pos. deviation from others' average contribution (0-20 Token) indicates how many Token the subject has contributed above the average of the other three group members, 0 if contributed less. Counter-punisher's neg. deviation from others' average contribution (0-20 Token, always non-negative) defined as how many Token the subject has contributed below the average of other three group members, 0 if contributed more. Received punishment (1-10 punishment points) is the number of received punishment points from the target in this period. Period (1-15) is the current period in this part of the experiment. FE location and treatment order means that a fixed effect for the location University of Zurich and a fixed effect for first treatment in session was included.

4 Conclusion

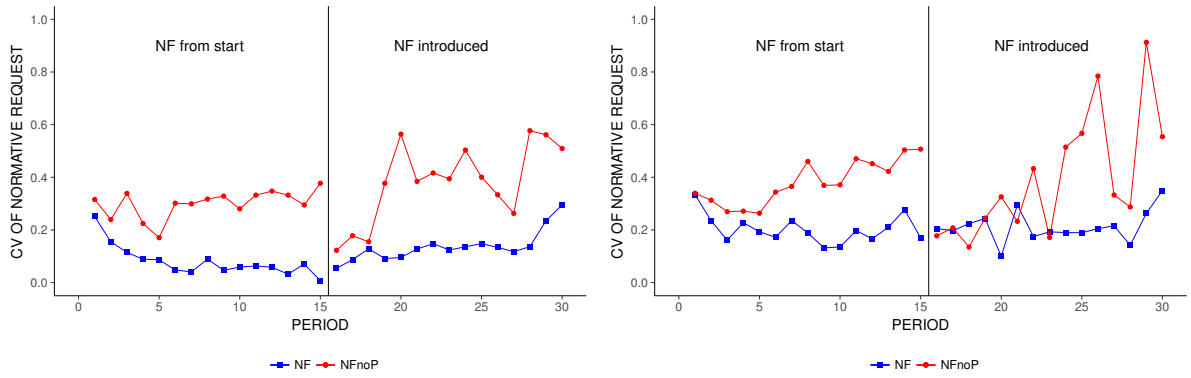
This paper experimentally examines two major questions regarding social norms. First, how and under what conditions social norms are formed and maintained and under what conditions they decay. Second, what the causal effects of social norms on behavior are,

specifically on cooperation and punishment behavior. We address both questions with the introduction of a simple norm formation opportunity to a laboratory public goods game. This allows studying several key properties of social norms empirically, namely their content, their strength and their stability. When means of enforcement exist, we observe the formation of strong and stable norms demanding contributions close to the surplus maximizing level. Without such means of enforcement, the social norm quickly decays. Social norm prove to be an effective and efficient mechanism to foster high contributions, but only when individuals had the possibility to sanction norm violators. This shows that whether or not social norms are merely cheap talk crucially depends on the ability to enforce them. Finally, we show that peer-punishment increases welfare when subjects are allowed to form social norms, contrary the results obtained when norm formation is ruled out.

5 Appendix Chapter IV

Additional Analyses

These additional analyses show that many of our result hold for our two experimental locations, Zurich and Nottingham, separately. Furthermore, we distinguish here between within-subject and between-subject analysis. Our subjects always participated in two treatments each repeated for 15 periods. We refer to between-subject comparisons when we only compare data from the first treatment alone. In within-subject comparisons we analyze changes when the first treatment does not feature the norm formation opportunity, but the second treatment does. Hence, these comparisons are based on data from sessions with the following order of treatments: either noNF–NF or noNFnoP–NFnoP. When a graph depicting periods from 1-30, period 1 up to 15 naturally refer to periods 1-15 of the first treatment and periods 16-30 refer to periods 1-15 of the second period. Empirical test are, unless otherwise stated, based on the pooled data from Zurich and Nottingham.

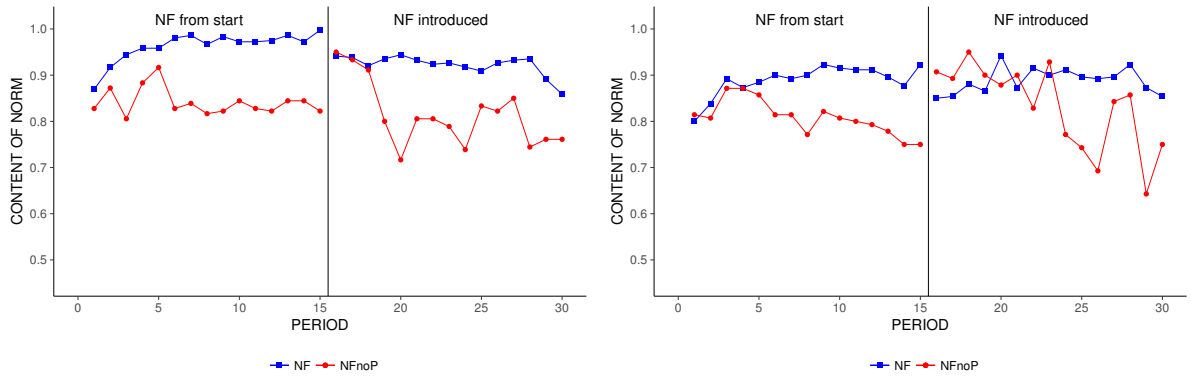


(a) Zurich (b) Nottingham

Figure 18: Coefficient of variation in normative requests over time indicating the strength of norm

Figure 18 shows how the coefficient of variation[26] (CV) evolves over time in (a) Zurich and (b) Nottingham separately. 'NF from start' means that the norm formation was present in the first treatment, and 'NF introduced' means that there is norm formation in the second treatment, but not in the first.

[26]Same definition as in main text.



(a) Zurich (b) Nottingham

Figure 19: Content and stability of norm

Note: Norms are normalized to 1, i.e. 1.0 corresponds to a requested contribution of the maximal possible level of 20 Token.

Figure 19 shows the evolution of the average content over time for each treatment that features the norm formation device for Zurich (a) and Nottingham (b) separately. In the first period, there is no significant difference in the content adopted by groups that have the possibility to punish compared to those that do not (Wilcoxon rank-sum test, n = 47 groups, p = 0.6906).[27] This also holds true for the first period with norm formation when it is newly introduced in period 16 (Wilcoxon rank-sum test, n = 46 groups, p = 1.000).[28] However, over the periods requested contributions in NF are slightly (+0.0021 per period) increasing, however, this increase is not significant (p = 0.1207).[29] On the contrary, without punishment, the content is significantly declining over time (-0.0079 per period; p = 0.0007).[30] This leads to significant differences when aggregating over all 15 periods with norm formation opportunities; requested contributions are higher both when norm formation is possible in the first treatment of the experiment (periods 1-15, Wilcoxon rank-sum test, n = 47 groups, p = 0.0021) and the second treatment (periods 16-30, Wilcoxon rank-sum test, n = 46 groups, p = 0.0653) when punishment is possible.[31]

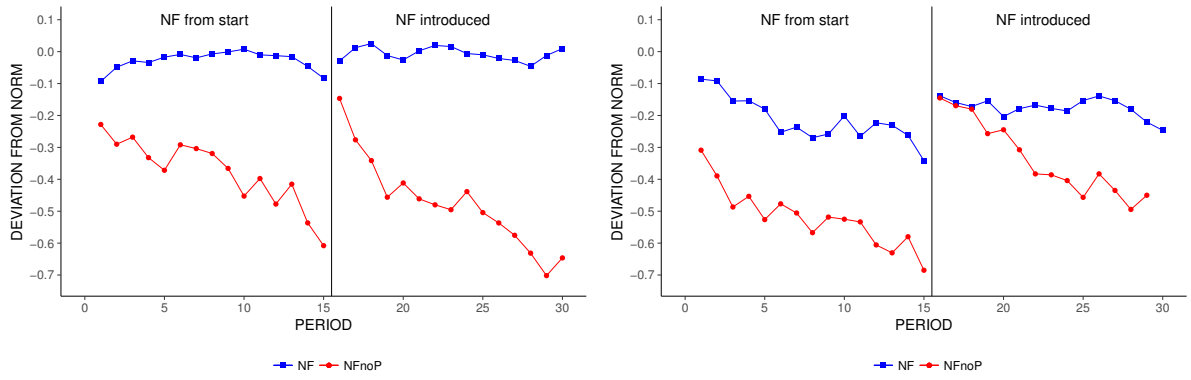
[27] Insignificant first period differences between NF and NFnoP are also present in an OLS regression of contributions on the treatment dummy (p = 0.6652), with fixed effect for location and clustered standard errors on the group level.

[28] Insignificant first period differences between NF and NFnoP are also present in an OLS regression of contributions on the treatment dummy (p = 0.3432), with fixed effect for location and clustered standard errors on the group level.

[29] These results are based on an OLS regression of groups' norms on periods in treatment condition NF with fixed effect for location and clustered standard errors on group level.

[30] These results are based on an OLS regression of groups' norms on periods in treatment NF with fixed effect for location and clustered standard errors on group level.

[31] Significant differences between NF and NFnoP are also present in OLS regressions of norm on treatment dummies (p = 0.0004 for periods 1-15 and p= 0.0570 for periods 16-30), with fixed effect for location and clustered standard errors on group level.

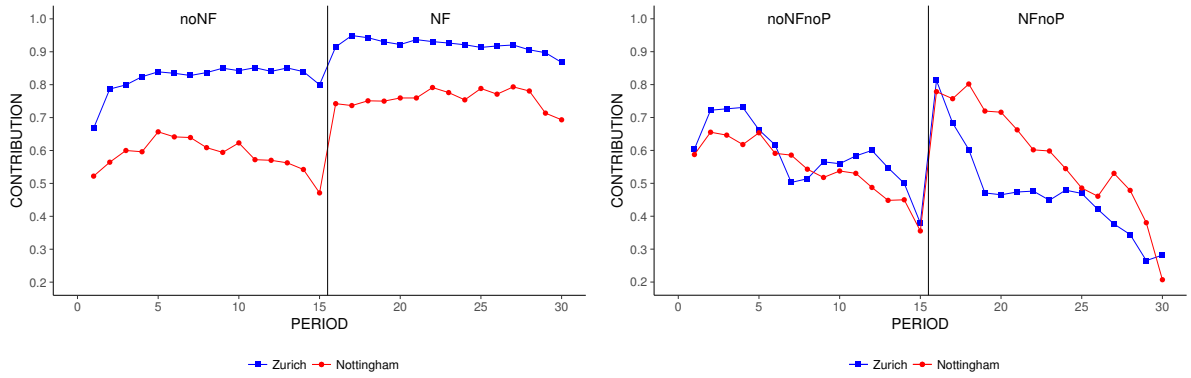


(a) Zurich (b) Nottingham

Figure 20: Deviations from norm over time

Note: Deviations from norm are normalized to 1. A deviation of 0 corresponds to a contribution that corresponds to the actually requested contribution, a -0.5 represents a contribution that is half of what was requested, and -1 indicates a contribution of 0 Token.

Figure 20 depicts the evolution of subjects' deviations in contributions, that is, actual minus by group requested contributions. 'NF from start' means that the norm formation was present in the first treatment, and 'NF introduced' means that there is norm formation in the second treatment, but not in the first.



(a) With punishment (b) Without punishment

Figure 21: Contributions when norm coordination introduced

Note: Contributions are normalized to 1, i.e. 1.0 corresponds to a contribution of the maximal possible level of 20 Token. Sub-figures (a) and (b) based on data from sessions noNF–NF and noNFnoP–NFnoP respectively (within-subject).

Figure 21 illustrates the change in contribution levels when subjects first participate in a treatment without norm formation and afterwards in one with such an opportunity (within-subject comparison). Trajectories for the case with punishment (sessions noNF–

NF) are depicted in sub-figure (a), and those without punishment (sessions noNFnoP–NFnoP) in sub-figure (b).

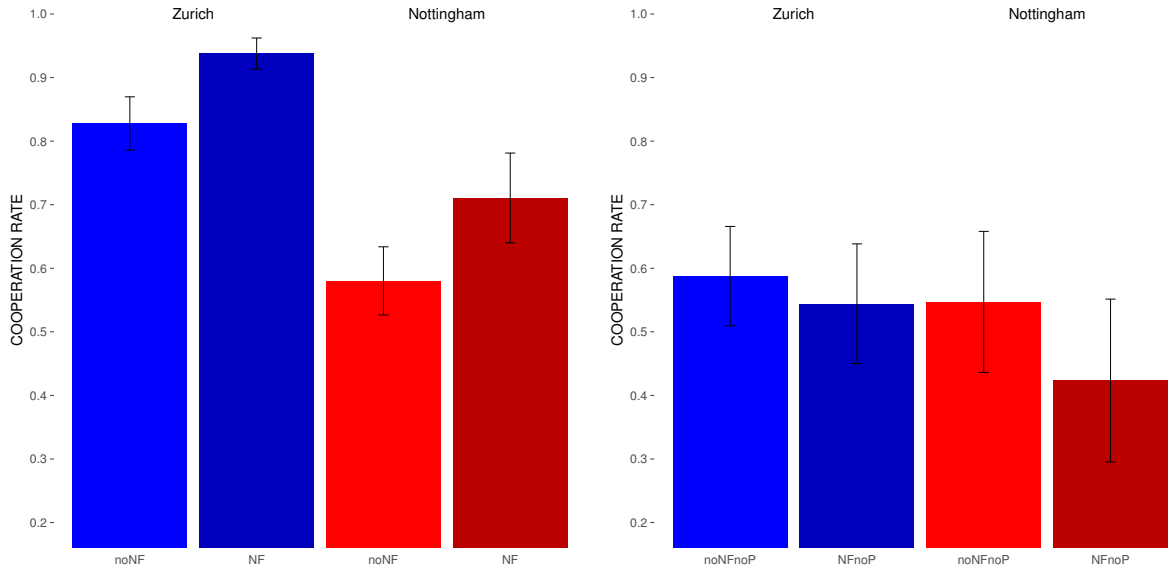
With punishment, the figure indicates that both in Zurich and in Nottingham the norm formation opportunity causes sizeable increases in average cooperation rates and the differences between NF and noNF remain fairly stable over time. In addition, the figure also shows that the cooperation enhancing role of the formation opportunity becomes effective immediately after its introduction: if one compares the first round of noNF with the first round of NF, one observes an increase in cooperation rates of roughly 20 percentage points. All these results are statistically significant. A Wilcoxon signed-rank test with groups' average contributions as independent observations shows that contributions in NF are significantly higher than in noNF (n = 30 groups, p = 0.0008).[32] The same test with groups' average contributions in period 1 of each treatment as the independent unit of observation yields p = 0.0000.[33]

This pattern contrast sharply with the pattern of cooperation in the absence of a punishment opportunity. Figure 21(b) shows that—although there is a sizable restart effect when the norm formation opportunity is introduced in period 16—cooperation rates quickly and strongly unravel with the norm coordination opportunity without punishment (NFnoP). In addition, the final cooperation levels in NFnoP (in period 30) are even lower than the final cooperation levels in noNFnoP (in period 15). Average contributions after the introduction are in fact smaller than before, however, not significantly so (Wilcoxon singed-rank test, n = 16 groups, p = 0.2934).[34]

[32]Significant differences between NF and noNF are also present in an OLS regression of contributions on the treatment dummy (p = 0.0015), with fixed effect for location and clustered standard errors on group level.

[33]Significant first period differences between NF and noNF are also present in an OLS regression of contributions on the treatment dummy (p = 0.0000), with fixed effect for location and clustered standard errors on group level.

[34]Insignificant negative differences between NFnoP and noNFnoP are also present in an OLS regression of contributions on the treatment dummy (p = 0.2811), with fixed effect for location and clustered standard errors on group level.



(a) With punishment (b) Without punishment

Figure 22: Average contribution (between-subject comparison)

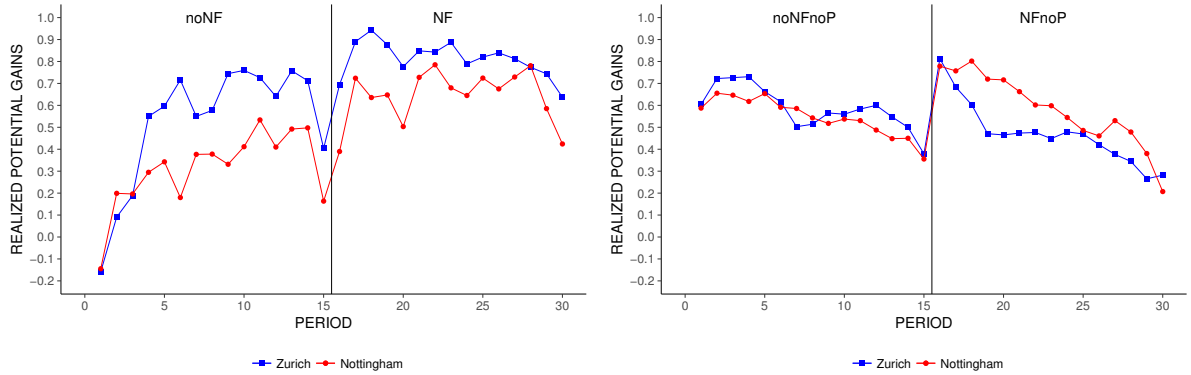
Note: Contributions are normalized to 1, i.e. 1.0 corresponds to a contribution of the maximal possible level of 20 Token. Sub-figures (a) and (b) based on data from initial treatment condition (between-subject). Bars depict clustered standard errors.

One may object that the comparison between NF and noNF may be influenced by the fact that NF is conducted after the experience of 15 periods of noNF. For this reason, we also conduct experiments in which treatment NF is conducted first which enables us to compare NF and noNF when subjects have no prior experience with either condition (between-subject comparison). This comparison is illustrated in figure 22(a), and it shows again that the NF condition leads to higher cooperation rates—a difference that is significant according to a Wilcoxon rank-sum test with group averages as units of observation (n = 84 groups, p = 0.0065).[35] Likewise, this significant difference between the NF and the noNF condition is already present in the first period (Wilcoxon rank-sum test, n = 84 groups, p = 0.0005).[36] Thus, when punishment is possible the salient social norms cause higher cooperation rates regardless of whether we compare treatments across time or at the beginning. For the case without punishment (comparing noNFnoP

[35]Significant differences between NF and noNF are also present in an OLS regression of contributions on the treatment dummy (p = 0.0113), with fixed effect for location and clustered standard errors on group level.

[36]Significant first period differences between NF and noNF are also present in an OLS regression of contributions on the treatment dummy (p = 0.0003), with fixed effect for location and clustered standard errors on the group level for the NF while the cluster for noNF is at the individual level. This difference in clustering is justified because in the first period of the noNF each individual contribution level constitutes and independent observation whereas in the first period of the NF subjects within a group may also have influenced each other because of the previous norm coordination stage in the game.

with NFnoP), figure 22(b) shows that a similar pattern as in the within-subject analysis emerges when we compare these two treatments when both are conducted at the beginning of an experimental session, that is, when subjects have not participated in a previous treatment; in this case the norm coordination opportunity also leads to slightly lower levels of cooperation although this decrease is not significant (Wilcoxon rank-sum test, n = 32 groups, p = 0.3860).[37]



(a) With punishment (b) Without punishment

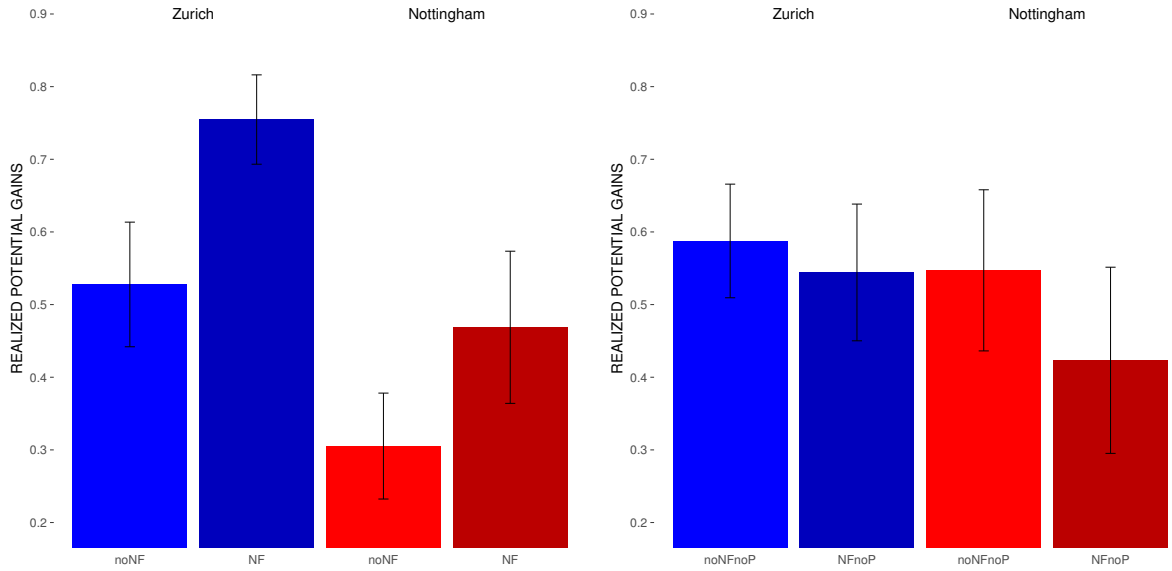
Figure 23: Group welfare when norm coordination introduced

Note: Realized Potential Gains are normalized to 1, i.e. 1.0 corresponds to a realized gain of the maximal possible level of 12 Token. Sub-figures (a) and (b) based on data from sessions noNF–NF and noNFnoP–NFnoP respectively (within-subject).

Figure 23(a) shows how the fraction of realized potential gains evolves over time when norm formation is introduced to an environment with punishment (within-subject comparison in sessions noNF–NF). Figure 23(b) shows the same trajectories when no punishment opportunities exist (sessions noNFnoP–NFnoP). Note that in NFnoP and noNFnoP realized potential gains are just a transformation of contributions, hence, the results about treatments differences are the same as for contributions, this is, no increase in efficiency without punishment. Regarding treatments with punishment, a Wilcoxon sign-rank test (n = 30 groups) relying on group averages per treatment reveals that the difference is statistically significant comparing either the first period in a treatment (p = 0.0000) or aggregating over the whole 15 periods in the treatment (p = 0.0000).[38]

[37] Insignificant negative differences between NF and noNF are also present in an OLS regression of contributions on the treatment dummy (p = 0.4147), with fixed effect for location and clustered standard errors on group level.

[38] Significant differences between NF and noNF are also present in an OLS regression of earnings on the treatment dummy first period alone (p = 0.0000) and all periods (p = 0.0000), with fixed effect for location and clustered standard errors on group level.



(a) With punishment (b) Without punishment

Figure 24: Average group welfare (between-subject comparison)

Note: Realized Potential Gains are normalized to 1, i.e. 1.0 corresponds to a realized gain of the maximal possible level of 12 Token. Sub-figures (a) and (b) based on data from initial treatment condition (between-subject). Bars depict clustered standard errors.

Again, one can object that our comparison is influenced by the order of treatments, therefore we analyze group welfare also on between-subjects basis. Figure 24 shows averages of realized potential gains from initial treatment conditions. With punishment, letting subjects first form a contribution norm boosts the fraction of realized gains in the first period on average by 29 p.p. (3.5 Token per peer, Wilcoxon rank-sum test, n = 84 groups, p = 0.0467).[39] Realized gains remain greater throughout the first 15 periods and are on average 20 p.p. above those under no norm coordination (Wilcoxon rank-sum test, n = 84 groups, p = 0.0183).[40]

[39] Significant differences between NF and noNF are also present in an OLS regression of earnings on the treatment dummy (p = 0.0852), with fixed effect for location and clustered standard errors on the group level for the NF while the cluster for noNF is at the individual level. This difference in clustering is justified because in the first period of the noNF each individual contribution level constitutes and independent observation whereas in the first period of the NF subjects within a group may also have influenced each other because of the previous norm coordination stage in the game.

[40] Significant differences between NF and noNF are also present in an OLS regression of earnings on the treatment dummy (p = 0.0467), with fixed effect for location and clustered standard errors on the group level.

Covariates

Table 7: Variables Regression Contribution

Variable	Meaning	Possible Values
Norm	Number of Token requested in this period as seen by subjects under NF	$0-20$
Contribution [t-1]	Contribution in previous period	$0-20$
Others' average contribution [t-1]	Average contribution of other group member in the previous period	$0-20$
Punishment [t-1]	Total number of received punishment points in the previous period	$0-30$
Period	Number of periods played	$1-15$
Zurich	Dummy variable for experiments in Zurich	No $(=0)$ or Yes $(=1)$
X * Y	Interaction between variable X and Y	

Table 8: Variables Regression Punishment

Variable	Meaning	Possible Values
Norm	Number of Token requested in this period as seen by subjects under norm coordination	$0-20)$
Average contribution w/o target	Average contribution of the group members excluding the potential target of punishment	$0-20$
Target's pos. deviation from norm	How much the target contributed more than requested by the norm	$0-20$
Target's pos. deviation from group	How much the target contributed more than the average of others	$0-20$
Target's neg. deviation from norm	How much the target contributed less than requested by the norm	$0-20$
Target's neg. deviation from group	How much the target contributed less than the average of others	$0-20$
Assigned punishment [t-1]	Total number of punishment points assigned by punisher in previous period	$0-10$
Rec. counter-punishment [t-1]	Total number of counter-punishment points received by punisher in previous period	$0-15$
Period this treatment	Current period in this treatment	$2-15$

Table 9: Variables Regression Counter-Punishment

Variable	Meaning	Possible Values
Pos. deviation norm	Counter-punisher's positive deviation in actual and requested contributions	$0-20$
Neg. deviation norm	Counter-punisher's negative deviation in actual and requested contributions	$0-20$
Pos. Deviation	How much the counter-punisher contributed more than the average of the rest of the group	$0-20$
Neg. Deviation	How much the counter-punisher contributed less than the average of the rest of the group	$0-20$
Received punishment	Number of punishment points received by counter-punisher from the target	$0-10$
Period this treatment	Current period in this treatment	$1-15$

Chapter V

Persuasion and Dissuasion in Immoral Labor Markets

Persuasion and Dissuasion in Immoral Labor Markets

Florian H. Schneider, Martin Schonger & Ivo Schurtenberger

Abstract

We study the supply of labor for immoral jobs, its relation to normative views, and to what extent labor supply and normative views can be shifted using persuasion and dissuasion. In the experiment, subjects are given the choice to perform a job which assists the marketing of tobacco products to young adults. Behavior is highly polarized: A quarter of subjects accepts the job for any positive wage, while another quarter of subjects refuses to do the 5-minute job for even $25. Attempts to persuade or dissuade subjects from working in the tobacco industry, created by a major tobacco company and the American Cancer Society respectively, do not shift labor supply. This finding can be explained by the firm normative views subjects hold.

JEL classification: C91; E24; A13

Keywords: Immoral labor markets; Corporate Image; Persuasion; Dissuasion; Moral Suasion; Tobacco; Social image

Citation: Schneider, F.H., Schonger, M., & Schurtenberger, I. (2018). Moral persuasion and dissuasion in immoral labor markets. *Working Paper*.

---

1 Introduction

Many industries exhibit or are perceived to exhibit large negative externalities. Prominent examples are the tobacco, arms and gambling industries. Take tobacco as a case in point: consumption of tobacco annually causes a loss of hundreds of thousands of quality-adjusted life years in the US alone (Kaplan et al., 2007). Working for such industries is often perceived to be immoral[41] (Frank, 1996; Brun et al., 2017), and consequently many employees suffer from a loss of purpose, reduced happiness, doubts about career choice and a bad social-image (Rosenblatt, 1994; Dolphin, 2005; Ashraf & Bandiera, 2017; Dur & van Lent, 2018). These psychic costs effect firms operating in immoral industries; current and prospective employees will require financial compensation for an immoral job, or even refuse to do it, thereby depressing labor supply and leading to higher equilibrium wages (Frank, 1996; Benedict et al., 2006; Brun et al., 2017).[42] Major companies deem difficulties in recruitment due to the immoral industry image as an important enough risk factor to warrant disclosure to regulators and shareholders Philip Morris International Inc. (2015); British American Tobacco (2015). Persuading employees that the firm is not immoral and that working for it is morally acceptable could mitigate these human resource problems. We study how susceptible potential employees are to such persuasion efforts by firms. Governments and civil society actors may want to shape the industry image in the opposite direction, thereby trying to dissuade prospective employees from working in immoral industries. Therefore, we also examine dissuasion efforts, where an industry is painted as immoral.

Various efforts to improve corporate image by immoral firms have been documented. For instance, Müller & Kräussel (2011) and Kotchen & Moon (2012) show that immoral firms invest more in corporate social responsibility (CSR) than non-immoral firms. Delmas & Burbano (2011) provide evidence that firms employing environmentally harmful business practices use "greenwashing," i.e. mislead the public about their behavior. In principle, such efforts can target consumers, employees, or regulators. For the case of the tobacco industry, internal company documents disclosed by court order (World Health Organization, 2004) show that tobacco companies have internal and external marketing programs with an express purpose of improving employee morale and recruitment (British American Tobacco, 1998; Philip Morris International Inc., 1999).

[41] For brevity, we refer to industries and jobs that are often perceived to be immoral as "immoral industries" and "immoral jobs." We do not intend to make a moral judgment thereby.

[42] Another consequence of these psychic costs can be decreased employee effort (Ariely et al., 2009; Carpenter & Gong, 2016). Relatedly, effort provision also decreases if there is no purpose in work (Ariely et al., 2008) or the employer expresses political views the employee does not like (Burbano, 2016).

These internal marketing programs furnish narratives about working for the company and the morality of doing so.[43] Narratives can serve as excuses for immoral behavior (Bénabou et al., 2018). Employees in immoral industries have an incentive to believe in particular narratives to decrease their psychic costs. A considerable body of evidence in psychology and economics, documents that people indeed tend to form self-serving beliefs (for a recent review, see Gino et al., 2016). This is reflected in the admission of a former tobacco trial lawyer "That's how you make a living, by rationalizing that black is not black, it's white, it's green, it's yellow" (Rosenblatt, 1994).

Corporations active in immoral industries are not the only ones trying to shape their image, so do their opponents (Spar & La Mure, 2003). Sometimes such efforts specifically target employees. For instance, at the Undersea Defense Technology conference in Glasgow 2018, company delegates had to walk past signs saying "global corporations responsible for arming the worst human rights abusers," and follow signs reading "death merchants this way" (BBC, 2018). Government actors also employ dissuasion efforts, for instance a U.S. federal court ordered four major tobacco companies to air the following statement "cigarette companies intentionally designed cigarettes with enough nicotine to create and sustain addiction" (Kessler, 2004).

The question arises how effective efforts to use persuasion and dissuasion to impact labor supply are. A challenge in investigating this question is that persuasion and dissuasion efforts often occur endogenously and concurrently, as companies and their opponents respond to political or social events or accidents,[44] which makes identification difficult. This paper uses a lab experiment to address this identification challenge.

In our experiment, participants are offered a real-effort job, which, as our data shows, many participants regard as socially inappropriate and immoral. The job aids the marketing of cigarettes to young adults. We elicit subjects' reservation wages for this job. Prior to elicitation of the reservation wage, study participants are randomly exposed to either a persuasion or dissuasion effort in relation to the job, or assigned to a control group. The treatment interventions used in the experiment are actual persuasion and dissuasion efforts employed by leading actors in the debate surrounding tobacco. In both cases they are high-quality, publicly available videos which they feature prominently in

[43] Two examples of narratives in the tobacco industry taken from a recruitment booklet are that "there is nothing in cigarettes that removes the ability of someone to stop smoking," and that adults have a fundamental right to make their own fully informed consumption choices (British American Tobacco, 1999).

[44] For instance, in the wake of the Deepwater Horizon oil spill, BP launched a corporate image campaign (Cherry & Sneirson, 2010), while Greenpeace activists scaled the BP headquarters to fly a flag depicting the oil stained BP logo (Guardian, 2010).

their online presences. The dissuasion effort is a video jointly produced by the American Cancer Society and the World Lung Foundation. It shows the suffering caused by tobacco and attacks the tobacco industry for its business practices. The persuasion intervention is the official company video of the large tobacco corporation, which manufactures the cigarettes used in the job. The video targets prospective white-collar employees. It highlights corporate social responsibility initiatives, the livelihoods of small farmers depending on tobacco, and the efforts the tobacco company undertakes to reduce risks and dangers associated with the consumption of its products. We choose these interventions as two large, well-funded organizations in the tobacco debate feature them prominently and thus must believe them to be particularly effective. We investigate the effect of the treatments on labor supply in the experiment, participants' willingness to work for the tobacco company (outside of the lab), their normative view about working for the tobacco company and their beliefs about the social appropriateness of accepting the job in the experiment.

We have three main results: The first result is that labor supply is highly polarized. 28% of subjects are willing to accept the job for CHF 1 ($\approx$ USD 1), the lowest possible (non-zero) reservation wage elicited. On the other hand, 25% of subjects decline to do the job for twenty-five times the pay, CHF 25, which was the maximum wage on the list. The polarization of reservation wages reflects subjects' heterogeneous views on the morality of working for tobacco and their heterogeneous beliefs about the social appropriateness of accepting the job.

The second result is that neither the persuasion nor the dissuasion effort has a statistically significant effect on reservation wages. Substantially, the coefficient estimates themselves are small. The study has 80% power to detect an effect of size Cohen's d of 0.5 at the 5% level. Also, the treatments have no statistically significant effect (at the 5%-level) on the stated willingness to accept employment with the tobacco company.

The third result sheds light on the reasons for these null-results: Our measures of normative view and social appropriateness indicate that participants have firm preexisting moral perceptions regarding aiding marketing of tobacco product. The company video neither affects the normative view of working for the tobacco company, nor the beliefs about the social appropriateness of doing the job. The video by the American Cancer Society has no statistically significant effect on the normative view of working for the tobacco company, but it does negatively affect the perceived social appropriateness of doing the job. However, the effect is small, which may explain why the video does not have detectable effect on labor supply.

Our second finding, that neither the persuasion nor the dissuasion intervention is effective, is intriguing in light of the fact that the two interventions we employ have been produced by sophisticated and well-funded organizations with decades of experience operating in a high-stakes controversy. The persuasion video is featured on the homepage of the tobacco company and as the top video on its YouTube channel. The American Cancer Society has a budget just short of one billion USD (American Cancer Society, 2017). Both videos are professionally produced, emotionally appealing and provide a plethora of arguments in favor of their positions. The experiment is designed to stack the odds in favor of finding an effect: subjects make a decision immediately after viewing the video, social image concerns are present, choices are highly incentivized—up to CHF 25 for a 5 minute job—and the job is closely related to the videos.

We contribute to an emerging literature that looks at the effect of corporate image improvements on labor supply. For studies that examine CSR as a recruitment tool, see Flammer & Luo (2017) and Bode et al. (2015). Cassar & Meier (2018) conduct a field experiment to investigate whether CSR can be deployed to increase employee effort. Firms donate to a charity to appear socially responsible. They find that if employees perceive the donations as instrumentally rather than intrinsically motivated they are ineffective. The question whether persuasion—another means to potentially improve corporate image—is effective has, to the best of our knowledge, not been explored. Our persuasion treatment provides evidence on this question: The null-results indicate that labor supply is irresponsive not only to instrumental CSR, but also to persuasion.

Our paper adds to the literature on strategic communication of moral excuses and narratives (Bénabou et al., 2018; Foerster & van der Weele, 2018a,b). The finding that people do not make use of the excuses provided to them in the persuasion treatment is in line with recent findings that show that even if experimental interventions provide excellent excuses or reasons not to follow norms, a large fraction of people stick to them. Van der Weele et al. (2014) find that people do not use "moral wiggle room" in the context of reciprocity. Ging-Jehli et al. (2018) demonstrate that individuals do not adopt negative beliefs about others' intentions in order to justify egoistic behavior. Bartling & Özdemir (2017) find that people do not employ the replacement logic ("if I don't do it, someone else will") in contexts with a strong social norm.

With the dissuasion treatment, we contribute to both the literature on reallocation of labor and the literature on moral suasion and moral reminders (Dal Bó & Dal Bó, 2014; Mazar et al., 2008; Verschuere et al., 2018). The market allocation of talent between industries with and without negative externalities is typically inefficient and it has been

suggested to use taxes to reallocate labor (Murphy et al., 1991; Mankiw, 2010; Rothschild & Scheuer, 2016; Lockwood et al., 2017). As an alternative to taxes, making use of moral suasion has been proposed (Pruckner & Sausgruber, 2013; Fellner et al., 2013; Luttmer & Singhal, 2014; Dwenger et al., 2016; Ito et al., 2018). The results of this study speak to the question to what extent moral suasion/dissuasion could be an alternative policy instrument for reallocation of labor. While we find that labor supply for immoral jobs reacts to monetary incentives, there is no evidence that moral suasion has an effect.

The remainder of the paper discusses describes the design of the experiment in section 2, section 3 describes the results, and section 4 concludes.

2 Study Design

For purposes of external validity, the immoral job and the persuasion efforts should meet three criteria: First, the job in the experiment should have consequences similar to those of an existing immoral job. Second, both persuasion and dissuasion efforts by sophisticated industry actors must be available and suitable for use in the lab. Third, working in the job should be socially observed, as is being employed by a particular employer or industry.

The Job

To meet these demands, we employ a novel but simple job. On each participant's desk there is a gift bag. The gift bag will be distributed to a young adult, regardless of the participant's choices. The participant has the option to gift-wrap three cigarettes and place them into the gift bag. Therefore, if the participant chooses to accept the job, she will cause a young adult to receive three cigarettes, otherwise, that young adult will not receive any cigarettes.[45] To make sure that subjects understand what the job entails, and what its consequences are, the following are placed on every subject's desk: the wrapping supplies (wrapping paper, stickers and ribbons), the gift bag, an example of a gift-wrapped cigarette, and a pack of cigarettes. The pack of cigarettes is open and contains exactly four cigarettes. The pack of cigarettes is an off-the-shelf pack of

[45]To ensure the study caused no harm, gift bags and cigarettes were only given to young adults who were regular smokers. From each gift bag recipient, we first bought four cigarettes. Thus, as a result of participation in the study, each smoker lost either one or four cigarettes. The study received authorization from the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich.

cigarettes with the standard warning text and images. All subjects receive cigarette packs with identical text and images. This immoral job has an industry analogue, working in the tobacco industry, and maybe more specifically in the marketing of tobacco.

Persuasion and Dissuasion Treatments

This job allows us to use real world persuasion and dissuasion efforts. In the persuasion treatment, the official company video of one of the world's leading tobacco companies is streamed to subjects' computers. This company is the manufacturer of the brand of cigarettes used in the job and subjects are made aware of that fact. The video highlights the positive effects the tobacco company has on its workforce, providing them with a purpose in life, and how they contribute to communities and society in general. The video features the stories of small tobacco farmers from around the globe, an office worker who finds purpose in his work, and scientists who research new and better products. The video shows that the company works closely together with local communities and provides them with crucial infrastructure. The video even touches upon the fact that as a new employee one might have concerns about working for a "controversial multinational." Substantively, the video makes the following points that could persuade our subjects to increase their labor supply by changing their normative perceptions: the company does not market to underage youth or children, it provides information and data to help its adult customers making informed choices, it conducts research to develop less harmful alternative products, small farmers and their families in poor countries benefit from cultivating tobacco, and the company invests in improving living conditions in the communities of these farmers. The video lasts four and a half minutes.

For the dissuasion effort, we use a video by what is presumably one of the world's leading foes of tobacco consumption, the American Cancer Society (produced jointly with the World Lung Foundation).[46] The video begins by asking how much a human life is worth. It then claims that for the tobacco industry it is $6000 as that is the profit made on average per person dying from smoking. The video states that tobacco will kill one in every three children who takes up smoking, and that smoking is a cause of cancer, heart diseases, lung disease, diabetes and more. According to the video, six million people die due to smoking every year. The video explains that smoking can be stopped by taxing cigarettes, passing smoking bans, by informing about the harms of smoking and prohibiting "slick advertising and packaging." The almost three-minute video concludes

[46]The video can be found on the YouTube channel of the American Cancer Society, https://www.youtube.com/watch?v=2m7-zIa6-Es.

by directly addressing the viewer, saying that "we are the solution," and asking her to "fight back," and telling her "you can help."

In the neutral control treatment, subjects watch a five-minute video that shows aerial footage of landscapes.

Labor Supply

After subjects have read all instructions and seen their respective video, we elicit their reservation wages for doing the job. To do so, we use the BDM mechanism (Becker et al., 1964). Specifically, we use the list method, where the list contains all integer amounts from CHF 0 to CHF 25. For each of these wages subjects choose whether they "accept" or "decline" the job. One wage is randomly drawn for the entire session. We do not impose a unique switching point. Occurrence of multiple switching points serves as an indicator of a subjects' failure to understand choice situation.

Social Image

Subjects' decisions to accept or decline the job are publicly observed: At the end of the session, subjects' decisions for the wage that was drawn and their pictures are displayed to all other participants in that session. The pictures are taken at the beginning of the session, so the fact that behavior is publicly observed is salient.

Moral Perception

We elicit two measures of moral perception. First, we ask how moral the participant thinks it is to work for the tobacco industry (*normative view*). We ask this indirectly: "Imagine one of your fellow students starts working for [name of tobacco company] after finishing his/her studies. How moral do you think this is?" This question is designed to capture participant's view on the morality of working for the tobacco company. Second, to measure participant's view on the social appropriateness of accepting the job (*social appropriateness*), we employ the method by (Krupka & Weber, 2013). Participants are asked how socially appropriate they think it is to accept and to decline the job for a wage of CHF 10.

Further Measures

To help assess the external validity of the experimental evidence, we ask subjects to indicate on a five-point Likert scale how willing they are to work for each in a list of 13 companies, two of which are tobacco companies, one of the latter being the manufacturer of the cigarettes used in the experiment.

Apart from asking how willing a subject is to work for the tobacco industry, we ask subjects to what extent any of twelve motives play a role in deciding when to accept or decline the job. The motives offered relate to effort, ability, consequences, positive and negative externalities, and morals. To learn about subjects' social preferences, we implement a trolley problem (Thomson, 1985) and a dictator game with an endowment of CHF 2.

Procedure

The sequence of the experiment is as follows: upon arrival participants have their pictures taken, and then enter the lab where they find detailed written instructions on their desks. They then answer ten understanding questions about the study. Immediately after being exposed to one of the three videos, they make their labor supply decisions. Subsequently, the additional measures are elicited, followed by a demographic questionnaire. Finally, the labor supply decisions of all participants are made public.

The experiment is programmed in z-Tree (Fischbacher, 2007), participants were recruited with hroot (Bock et al., 2014). Subjects were students from the joint subject pool of the University of Zurich and the Swiss Federal Institute of Technology. Only self-reported non-smokers participated in our study to have a more homogeneous subject pool. The sessions lasted about 60 minutes. Average earnings were about CHF 34.

3 Results

We conducted 18 experimental sessions with 10 to 13 participants per session. For each of the three treatments, six sessions were conducted, yielding 65 participants in the persuasion treatment, 66 in the neutral control treatment, and 73 in the dissuasion treatment, giving a total of 204 participants in the study. All but one subject exhibit a monotone willingness to accept the job, that is, only one subject switched between accepting and declining the job more than once. This individual is excluded from the analysis. The

fact that virtually all subjects exhibit a monotone willingness to accept can be seen as an indicator that subjects understood the labor supply choice they were facing.

Figure 25 depicts the cumulative distribution function of reservation wages pooled over all three treatments. Subjects who decline to do the job for any of the wages offered are grouped together (label >25). About 28% of subjects have a reservation wage that is equal to or less than CHF 1, the lowest (non-zero) wage offered. By contrast, about 25% of subjects decline the job for any of the wages offered, including the highest wage offered of CHF 25.

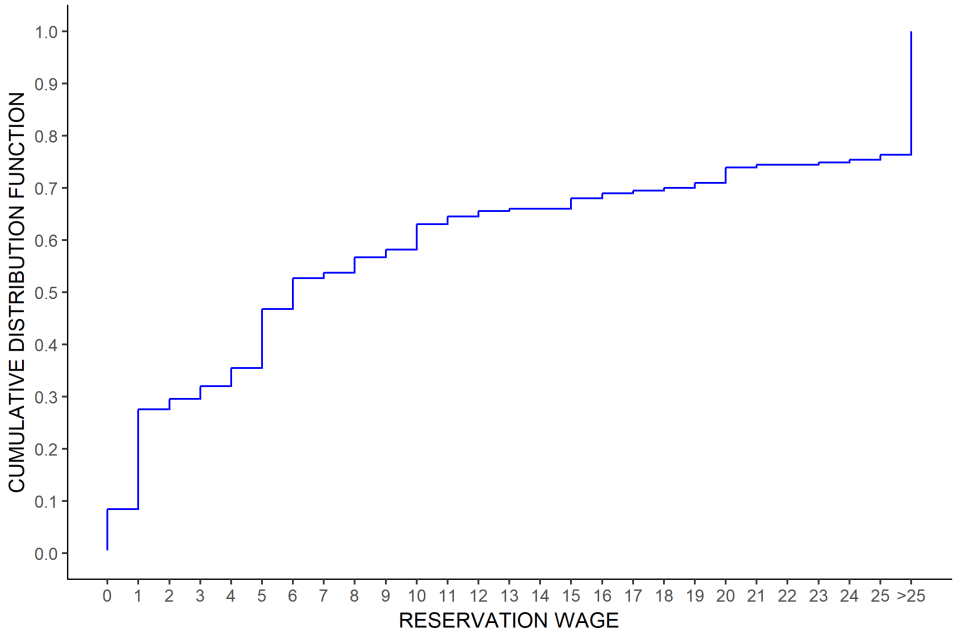


Figure 25: Cumulative distribution function of reservation wages

This highly polarized behavior warrants a closer look at subjects who decline to do the job for any wage. Almost all (94%) of these subjects, indicate that the argument "it is immoral to wrap the cigarettes" affected or greatly affected their decision. By contrast non-moral related reasons do not seem to play a role. No subject gives any of the three non-moral related, prewritten statements (effort, not knowing how to wrap the cigarettes, considering themselves bad at wrapping cigarettes) as a reason for her behavior. The single exception to that is one subject who indicates that effort mattered for her decision somewhat.

Across all subjects there is substantial disagreement on the morality of performing the job in the lab, as well as accepting jobs in the tobacco industry. While 45% of subjects have a negative normative view of accepting a job in the tobacco industry, 49% of subjects

have a neutral normative view. Furthermore, while 55% of all subjects believe that it is at least somewhat socially inappropriate to accept the job for a wage of 10 CHF, 30% of subjects think it is neither socially appropriate nor inappropriate. Figure 27 in the appendix of chapter V gives the full distributions.

Can these divergent views help explain the polarization of subjects' reservation wages? To look at this issue, we estimate a regression model of reservation wages[47] on normative views and believed social appropriateness. As our data is censored at CHF 0 and CHF 26, we specify a Tobit model. As can be seen in Table 10, the coefficient estimates for both variables are statistically significant. Hence, reservation wages seem to reflect the perceived immorality of accepting the job. The effect sizes are substantial: On average, a subject's reservation wage is estimated to be more than CHF 10 higher if she considers working for the cigarette manufacturer as "very immoral" compared to someone who regards this as morally "neutral" (normative view). Similarly, reservation wages are on average more than CHF 7 higher for subjects who believe it is "socially very inappropriate" to accept the job in the lab compared to those subjects who consider this as "neither socially inappropriate nor socially appropriate." These estimates are robust to adding individual controls (column 2).

[47]Note that our data is discrete due to use of the list method. Hence, technically, for a subject, who accepts the job for wage ω, but declines it for wage $\omega - 1$, the reservation wage must be in $[\omega - 1, \omega]$. We set the reservation wage to ω.

Table 10: Relationship between reservation wages and moral perception

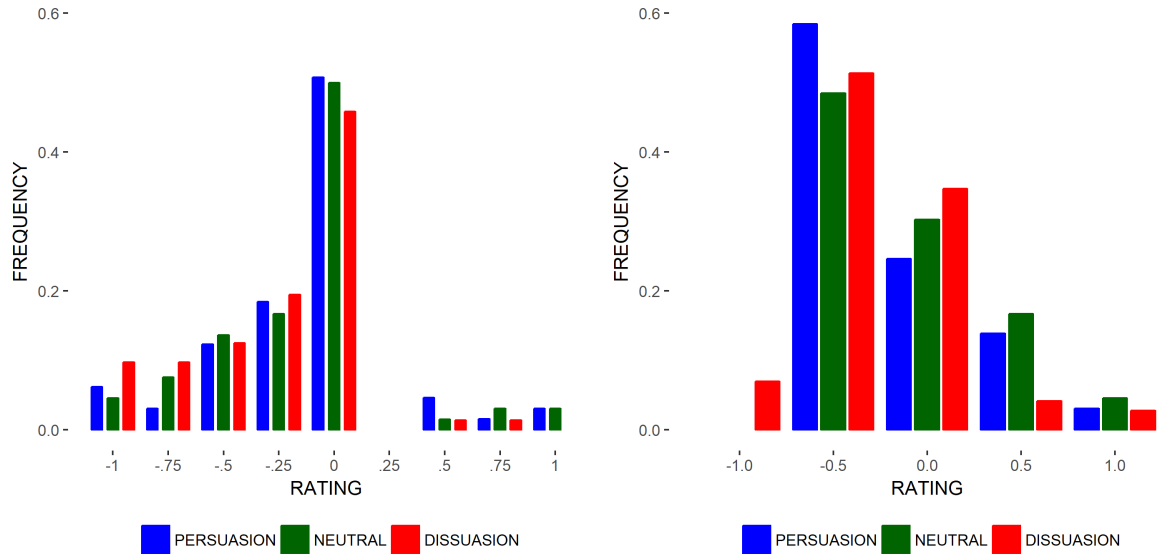
	Tobit Regression	
	Dependent variable: *reservation wage*	
	(1)	(2)
normative view	-10.26^{***}	-10.57^{***}
	(-3.85)	(-4.06)
social appropriateness	-7.10^{***}	-7.30^{***}
	(-2.99)	(-3.03)
constant	8.51^{***}	3.68
	(-7.92)	(-0.66)
log(sigma)	2.56^{***}	2.52^{***}
	(38.74)	(38.27)
n	203	203
log-likelihood	-612.30	-606.56
control variables	NO	YES

Notes: Coefficient estimates of Tobit regressions, left-censored at 0 (n=17), right-censored at 26 (n=48). Independent variables: normative view measures how moral the subject rates working for the manufacturer of the cigarettes (from −1 very immoral to 1 very moral), social appropriateness measures subjects' beliefs about how appropriate others view accepting the job (from −1 very socially inappropriate to 1 very socially appropriate). Control variables: gender, age, education, choice in trolley problem and dictator giving. t-statistics in parentheses; *p<0.05; **p<0.01; ***p<0.001.

These results show that it is the subjects who perceive accepting the job as immoral that demand a wage premium. The magnitude of the estimated coefficients suggests that successfully changing normative views and/or beliefs about the social appropriateness would be a promising avenue to alter labor supply in either direction.

The challenge organizations then face is how to alter normative views and beliefs about social appropriateness. Figure 26 shows how effective the persuasion and dissuasion efforts by leading actors are in the lab. Figure 26 (a) shows the relative frequencies of subjects' normative views. Individuals have firm normative views and notions of appropriate behavior: There is no statistically significant difference between the distribution in the neutral and in the persuasion treatment (Wilcoxon rank sum, p=0.665) nor between the distribution in the neutral and in the dissuasion treatment (rank sum, p=0.269). Figure 26 (b) depicts how appropriate subjects think others regard accepting the job. Again, there is no statistically significant difference between the distribution in the neutral and in the persuasion treatment (rank sum, p=0.269). By contrast, there is a statistically significant difference between the distribution in the neutral and in the dissuasion treat-

ment (rank sum, p=0.044). For the complementary Krupka-Weber question of declining (rather than accepting) the job, neither treatment has a statistically significant effect (see Figure 28 in the appendix of chapter V).



(a) Normative view (b) Social appropriateness of accepting the job

Figure 26: Appropriateness ratings across treatments

Notes: normative view measures how moral the subject rates working for the manufacturer of the cigarettes (from −1 very immoral to 1 very moral), social appropriateness measures subjects' beliefs about how appropriate others view accepting the job (from −1 very socially inappropriate to 1 very socially appropriate).

Based on these small and mostly insignificant effects on normative views and beliefs, a shift in reservation wages is not to be expected. Table 11 confirms this. It gives the estimates of a Tobit regression of reservation wages on treatment groups, where the neutral treatment is the omitted category. There is no statistically significant effect of the persuasion treatment (p=0.470), nor of the dissuasion treatment (p=0.534). The sample size gives us the power to reject substantial effect sizes on reservation wages: The statistical power with N=203 was 80% to detect an effect size equal to a Cohen's d of 0.50 at the 5% level (using powerBBK, developed by Bellemare et al., 2016). This result is robust to controlling for gender, age and education (column 2).

Table 11: Effect of treatments on reservation wages

	Tobit Regression	
	Dependent variable: *reservation wage*	
	(1)	(2)
persuasion treatment	1.83	1.79
	(0.72)	(0.71)
dissuasion treatment	1.54	1.67
	(0.62)	(0.68)
constant	10.74***	11.11*
	(6.05)	(1.82)
log(sigma)	2.64***	2.63***
	(39.89)	(39.71)
n	203	203
log-likelihood	−629.85	−626.90
control variables	NO	YES

Notes: Coefficient estimates of Tobit regressions, left-censored at 0 (n=17), right-censored at 26 (n=48). Control variables: gender, age and education. t-statistics in parentheses; *p<0.05; **p<0.01; ***p<0.001.

As we have seen, the interventions do not successfully change behavior in the lab. A potential concern in the interpretation of these results is that the job in the lab is different from employment for a tobacco company, and this may be why, in the lab, they are ineffective. To investigate this, we ask subjects how willing they are to accept a job offer at the tobacco company.[48] A Wilcoxon rank sum test finds no statistically significant difference in willingness to work for the tobacco company between the neutral and the persuasion treatment (p=0.674), respectively the dissuasion treatment (p=0.072). Figure 29 in the appendix of chapter V gives the distribution of willingness to work for all three treatments. Indeed, behavior in our laboratory job is closely mirrored in subjects' stated willingness to work for the tobacco company: a Tobit regression of reservation wages on stated willingness to work shows that on average a subject who stated to be "somewhat unwilling" to work for the tobacco company exhibits a more than CHF 10 higher reservation wage for the laboratory job than a subject who answered with "somewhat willing" (p=0.000).

[48] Wiswall & Zafar (2018) provide evidence that such stated preferences are predictive of ultimate employment.

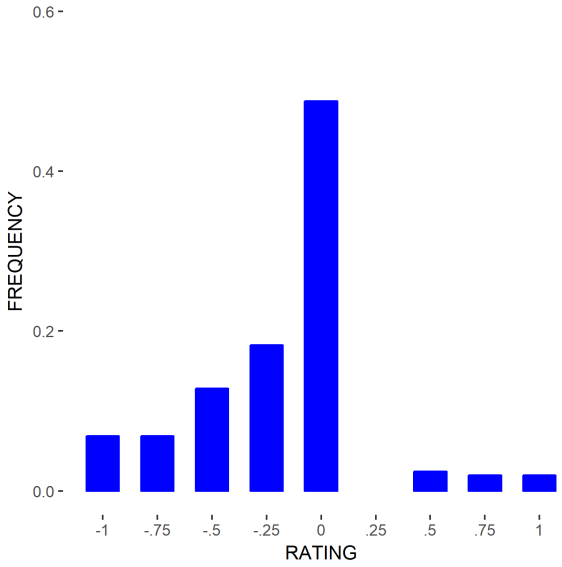
4 Conclusion

We study whether normative views regarding immoral jobs, and the labor supply for such jobs, can be shifted using professionally-made persuasion and dissuasion attempts. To do so, we employ a novel lab paradigm for studying immoral labor markets. The paradigm has two key strengths: first, the laboratory job has an external analogue (marketing of a harmful and addictive consumer product), and second, real-world persuasion and dissuasion attempts can be used in the lab.

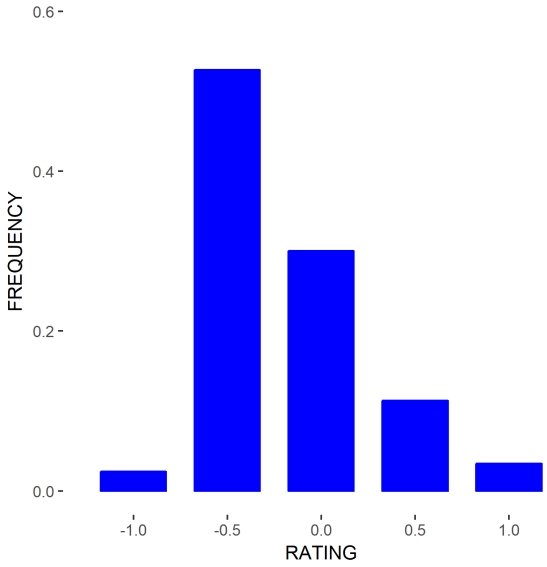
Subjects' behavior regarding the immoral job is highly polarized. About one quarter of subjects is willing to perform the job even for the lowest wage offered, in contrast another quarter of subjects refuses to do the job at the highest wage we offer them—$ 25 for five minutes of work. This heterogeneity in reservation wages can be explained by subjects' normative views about working for the tobacco company and their beliefs about the social appropriateness of accepting the job. Given the strong link between behavior and norms evident in the data, persuasion and dissuasion efforts have the potential to substantially alter labor supply. However, it turns out that both the persuasion and dissuasion efforts do not influence labor supply: Neither does the company video convince participants to accept the job, nor does the video by the American Cancer Society convince people to decline the job.

Our results suggest that, at least in the context of the immoral labor market at hand, individuals tend to have firm norms and notions of appropriate behavior. This largely limits the scope for such persuasion and dissuasion efforts.

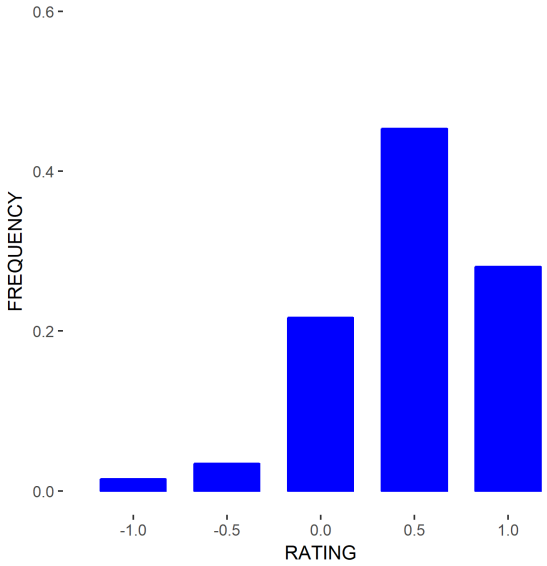
5 Appendix Chapter V



(a) Normative view



(b) Social appropriateness of accepting the job



(c) Social appropriateness of declining the job

Figure 27: Distribution normative perception

Notes: normative view measures how moral the subject rates working for the manufacturer of the cigarettes (from −1 very immoral to 1 very moral), social appropriateness of accepting/declining the job measures subjects' beliefs about how appropriate others view accepting the job (from −1 very socially inappropriate to 1 very socially appropriate).

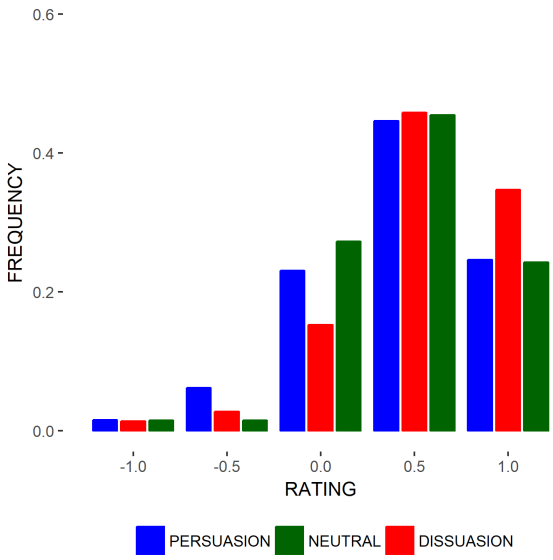


Figure 28: Social appropriateness of declining the job across treatments

Notes: Social appropriateness measures subjects' beliefs about how appropriate others view declining the job (from −1 very socially inappropriate to 1 very socially appropriate). There is no statistically significant difference between the distribution of the neutral and the persuasion treatment (ranksum, p=0.883) nor between the distribution of the neutral and the dissuasion treatment (ranksum, p=0.107).

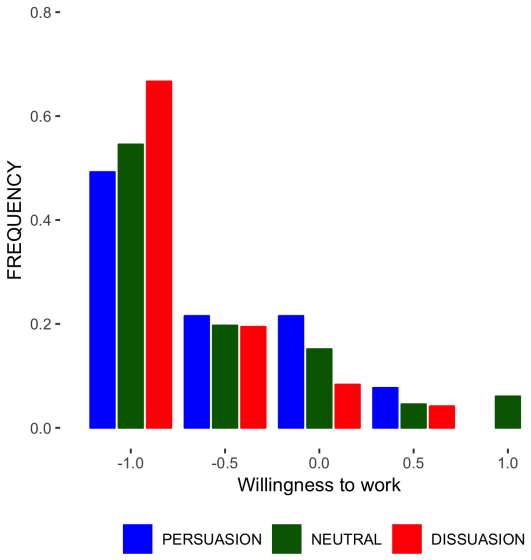


Figure 29: Willingness to work for the tobacco firm

Notes: Willingness to work measures subjects' willingness to work for the tobacco company that produced the company video (from −1 Not willing at all to 1 Very much willing).

References

References

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *Quarterly Journal of Economics* 94(4), 749–775.

Akerlof, G. A. (2007). The missing motivation in macroeconomics. *American Economic Review* 97(1), 5–36.

Alesina, A., Giuliano, P., & Nunn, N. (2013). On the origins of gender roles: women and the plough. *Quarterly Journal of Economics* 128(2), 469–530.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics* 95(9-10), 1082–1095.

Ambrus, A., & Greiner, B. (2012). Imperfect public monitoring with costly punishment: An experimental study. *The American Economic Review* 102(7), 3317–3332.

Ambrus, A., & Greiner, B. (2015). Democratic punishment in public good games with perfect and imperfect observability. *Economic Research Initiatives at Duke (ERID) Working Paper* (183).

Ambrus, A., & Pathak, P. A. (2011). Cooperation over finite horizons: a theory and experiments. *Journal of Public Economics* 95(7-8), 500–512.

American Cancer Society (2017). *Annual Stewardship Report*. Report. URL: https://www.cancer.org/content/dam/cancer-org/online-documents/en/pdf/reports/2017-annual-stewardship-report.pdf.

Andreoni, J. (1988). Why free ride? strategies and learning in public goods experiments. *Journal of Public Economics* 37(3), 291–304.

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110(1), 1–21.

Andreoni, J., Erard, B., & Feinstein, J. (1998). Tax compliance. *Journal of economic literature* 36(2), 818–860.

Andreoni, J., & Gee, L. K. (2012). Gun for hire: delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics* 96(11), 1036–1046.

Aoyagi, M., & Fréchette, G. (2009). Collusion as public monitoring becomes noisy: Experimental evidence. *Journal of Economic Theory* 144(3), 1135–1165.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review* 99(1), 544–555.

Ariely, D., Kamenica, E., & Prelec, D. (2008). Man's search for meaning: The case of legos. *Journal of Economic Behavior & Organization* 67(3-4), 671–677.

Ashraf, N., & Bandiera, O. (2017). Altruistic capital. *American Economic Review* 107(5), 70–75.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *Journal of conflict resolution* 24(1), 3–25.

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences* 108(27), 11023–11027.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological Bulletin* 137(4), 594–615.

Barr, A., Packard, T., & Serra, D. (2014). Participatory accountability and collective action: experimental evidence from albania. *European Economic Review* 68, 250–269.

Bartling, B., & Özdemir, Y. (2017). The limits to moral erosion in markets: Social norms and the replacement excuse. *Working Paper*.

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review* 97(2), 170–176.

BBC (2018). Two arrested at protests in glasgow to mark military fair. *BBC News*, 26 June.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science* 9(3), 226–232.

Bellemare, C., Bissonnette, L., & Kröger, S. (2016). Simulating power of economic experiments: the powerbbk package. *Journal of the Economic Science Association* 2(2), 157–168.

Bénabou, R., Falk, A., & Tirole, J. (2018). Narratives, imperatives and moral reasoning. *Working Paper*.

Benabou, R., & Tirole, J. (2011a). Identity, morals, and taboos: beliefs as assets. *Quarterly Journal of Economics* 126(2), 805–855.

Benabou, R., & Tirole, J. (2011b). Laws and norms. *Working Paper*.

Benedict, M. E., McClough, D., & McClough, A. C. (2006). The price of morals: An empirical investigation of industry sectors and perceptions of moral satisfaction—do business economists pay for morally satisfying employment? *The American Economist* 50(1), 21–36.

Berkowitz, L. (1972). Social norms, feelings, and other factors affecting helping and altruism. *Advances in Experimental Social Psychology* 6, 63–108.

Berkowitz, L., & Daniels, L. R. (1964). Affecting the salience of the social responsibility norm: effects of past help on the response to dependency relationships. *The Journal of Abnormal and Social Psychology* 68(3), 275–281.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy* 102(5), 841–877.

Bicchieri, C. (2002). Covenants without swords: group identity, norms, and communication in social dilemmas. *Rationality and Society* 14(2), 192–228.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2017). *Norms in the wild*. New York: Oxford University Press.

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization* 60(1), 11–26.

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review* 71, 117–120.

Bode, C., Singh, J., & Rogan, M. (2015). Corporate social initiatives and employee retention. *Organization Science* 26(6), 1702–1720.

Bolton, G. E., & Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review* 90(1), 166–193.

Bosman, R., Sutter, M., & van Winden, F. (2005). The impact of real effort and emotions in the power-to-take game. *Journal of Economic Psychology* 26(3), 407–429.

Botsford, L. W., Castilla, J. C., & Peterson, C. H. (1997). The management of fisheries and marine ecosystems. *Science* 277(5325), 509–515.

Boyd, R., & Richerson, P. (2005). Solving the puzzle of human cooperation. In S. Levinson, & P. Jaisson (Eds.), *Evolution and Culture*. Cambridge MA: MIT Press.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology* 13(3), 171–195.

Boyd, R., & Richerson, P. J. (1994). The evolution of norms - an anthropological view. *Journal of Institutional and Theoretical Economics* 150(1), 72–87.

Breza, E., Kaur, S., & Krishnaswamy, N. (2018). Scabs: norm-driven suppression of labor supply. *Working Paper*.

British American Tobacco (1998). *Corporate social responsibility. Note to Martin Broughton from Heather Honour*. Report. URL: https://www.industrydocumentslibrary.ucsf.edu/tobacco/docs/#id=pkgx0195.

British American Tobacco (1999). *Taling about tobacco*. Report. URL: https://www.industrydocumentslibrary.ucsf.edu/tobacco/docs/#id=spvp0202.

British American Tobacco (2015). *Annual Report*. Report. URL: http://www.bat.com/group/sites/uk__9d9kcy.nsf/vwPagesWebLive/DO9DCL3B/$FILE/medMDA87PVT.pdf?openelement.

Brosnan, S. F., Schiff, H. C., & De Waal, F. B. (2005). Tolerance for inequity may increase with social closeness in chimpanzees. *Proceedings of the Royal Society of London B: Biological Sciences* 272(1560), 253–258.

Brown, M., Falk, A., & Fehr, E. (2004). Relational contracts and the nature of market interactions. *Econometrica* 72(3), 747–780.

Bruhin, A., Fehr, E., & Schunk, D. (forthcoming). The many faces of human prosociality. *Journal of the European Economic Association*.

Brun, F., Weber, R. A., & Schneider, F. H. (2017). Immoral labor markets. *Working Paper*.

Burbano, V. C. (2016). Social responsibility messages and worker wage requirements: Field experimental evidence from online labor marketplaces. *Organization Science* 27(4), 1010–1028.

Burks, S., Nosenzo, D., Anderson, J., Bombyk, M., Ganzhorn, D., Götte, L., & Rustichini, A. (2016). Lab measures of other-regarding preferences can predict some related on-the-job behavior: evidence from a large scale field experiment. *Working Paper*.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83(2), 284–299.

Carlsson, F., Johansson-Stenman, O., & Nam, P. K. (2014). Social preferences are stable over long periods of time. *Journal of Public Economics* 117, 104–114.

Carpenter, J., & Gong, E. (2016). Motivating agents: How much does the mission matter? *Journal of Labor Economics* 34(1), 211–236.

Carpenter, J. P. (2007). Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior* 60(1), 31–51.

Carpenter, J. P., & Matthews, P. H. (2012). Norm enforcement: anger, indignation, or reciprocity? *Journal of the European Economic Association* 10(3), 555–572.

Cassar, L., & Meier, S. (2018). *Intentions for Doing Good Matter for Doing Well: The (Negative) Signaling Value of Prosocial Incentives*. Report National Bureau of Economic Research.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* 14(1), 47–83.

Chaudhuri, A., Graziano, S., & Maitra, P. (2006). Social learning and norms in a public goods experiment with inter-generational advice. *Review of Economic Studies* 73(2), 357–380.

Cherry, M. A., & Sneirson, J. F. (2010). Beyond profit: Rethinking corporate social responsibility and greenwashing after the bp oil disaster. *Tulane Law Review* 85, 983.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology* 24, 201–234.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics* 9(3), 265–279.

Conlin, M., Lynn, M., & O'Donoghue, T. (2003). The norm of restaurant tipping. *Journal of Economic Behavior & Organization* 52(3), 297–321.

Crawford, S. E., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review* 89(03), 582–600.

Croson, R. (1996). Partners and strangers revisited. *Economics Letters* 53(1), 25–32.

Cubitt, R. P., Drouvelis, M., Gächter, S., & Kabalin, R. (2011). Moral judgments in social dilemmas: how bad is free riding? *Journal of Public Economics* 95, 253–264.

Cummins, D. D. (1996). Evidence of deontic reasoning in 3-and 4-year-old children. *Memory & Cognition* 24(6), 823–829.

Dal Bó, E., & Dal Bó, P. (2014). "do the right thing:" the effects of moral suasion on cooperation. *Journal of Public Economics* 117, 28–38.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature* 446(7137), 794–796.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology* 31(1), 169–193.

Dawes, R. M., McTavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology* 35(1), 1–11.

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review* 54(1), 64–87.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33(1), 145–167.

DeQuervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science* 305(5688), 1254–1258.

Dhami, S., Wei, M., & Al-Nowaihi, A. (forthcoming). Public goods games and psychological utility: theory and evidence. *Journal of Economic Behavior & Organization*.

Dickson, E. S., Gordon, S. C., & Huber, G. A. (2009). Enforcement and compliance in an uncertain world: An experimental investigation. *The Journal of Politics* 71(04), 1357–1378.

Dolphin, R. R. (2005). Internal communications: Today's strategic imperative. *Journal of Marketing Communications* 11(3), 171–190.

Dufwenberg, M., Gachter, S., & Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior* 73(2), 459–478.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior* 47(2), 268–298.

Dur, R., & van Lent, M. (2018). Socially useless jobs. *Working Paper*.

Dwenger, N., Kleven, H., Rasul, I., & Rincke, J. (2016). Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in germany. *American Economic Journal: Economic Policy* 8(3), 203–32.

Ellickson, R. C. (2001). The evolution of social norms: a perspective from the legal academy. In M. Hechter, & K. D. Opp (Eds.), *Social Norms* (pp. 35–75). New York: Russell Sage Foundation.

Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2012). Social framing effects: preferences or beliefs? *Games and Economic Behavior* 76(1), 117–130.

Elster, J. (1989a). *The cement of society: A survey of social order*. Cambridge University Press.

Elster, J. (1989b). Social norms and economic theory. *Journal of Economic Perspectives* 3(4), 99–117.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? individual decisions and majority rule in mitigating the free rider problem. *European Economic Review* 53(5), 495–511.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior* 54(2), 293–315.

Feess, E., Schramm, M., & Wohlschlegel, A. (2014). The impact of fine size and uncertainty on punishment and deterrence: Evidence from the laboratory. *available at SSRN-id2464937*.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature* 425(6960), 785.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences* 8(4), 185–190.

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior* 25(2), 63–87.

Fehr, E., & Gächter, S. (2000a). Cooperation and punishment in public goods experiments. *The American Economic Review* 90(4), 980–994.

Fehr, E., & Gächter, S. (2000b). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 14(3), 159–181.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415(6868), 137–140.

Fehr, E., Hoff, K., & Kshetramade, M. (2008). Spite and development. *American Economic Review* 98(2), 494–499.

Fehr, E., & Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the tragedy of the commons. *Journal of Public Economics* 95(9-10), 1144–1155.

Fehr, E., & List, J. A. (2004). The hidden costs and returns of incentives-trust and trustworthiness among ceos. *Journal of the European Economic Association* 2(5), 743–771.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature* 422(6928), 137–140.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114(3), 817–868.

Fehr, E., & Schurtenberger, I. (2018a). The dynamics of norm formation and norm decay. *Working Paper, Department of Economics, University of Zurich*.

Fehr, E., & Schurtenberger, I. (2018b). Normative foundations of human cooperation. *Nature Human Behaviour* 2(7), 458.

Fehr, E., & Williams, T. (2018). Social norms, endogenous sorting and the culture of cooperation. *Working Paper*.

Fellner, G., Sausgruber, R., & Traxler, C. (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association* 11(3), 634–660.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2), 171–178.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review* 100(1), 541–556.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters* 71(3), 397–404.

Fischer, S., Grechenig, K. R., & Meier, N. (2013). Cooperation under punishment: Imperfect information destroys it and centralizing punishment does not help. *MPI Collective Goods Preprint* 6.

Flammer, C., & Luo, J. (2017). Corporate social responsibility as an employee governance tool: Evidence from a quasi-experiment. *Strategic Management Journal* 38(2), 163–183.

Foerster, M., & van der Weele, J. J. (2018a). Denial and alarmism in collective action problems. *Working Paper*.

Foerster, M., & van der Weele, J. J. (2018b). Persuasion, justification and the communication of social impact. *Working Paper*.

Frank, R. H. (1996). What prices the moral high ground? *Southern Economic Journal* 63(1), 1–17.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1518), 791–806.

Gächter, S., & Herrmann, B. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural russia. *European Economic Review* 55(2), 193–210.

Gächter, S., Nosenzo, D., & Sefton, M. (2013). Peer effects in pro-social behavior: social norms or social preferences? *Journal of the European Economic Association* 11(3), 548–573.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science* 322(5907), 1510–1510.

Gächter, S., & Thöni, C. (2005). Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association* 3(2-3), 303–314.

Gelcich, S., Guzman, R., Rodríguez-Sickert, C., Castilla, J. C., & Cárdenas, J. C. (2013). Exploring external validity of common pool resource experiments: insights from artisanal benthic fisheries in chile. *Ecology and Society* 18(3).

Ging-Jehli, N., Schneider, F. H., & Weber, R. A. (2018). On self-serving strategic beliefs. *Working Paper*.

Gino, F., Norton, M. I., & Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives* 30(3), 189–212.

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment despite reasonable doubt - a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies* 7(4), 847–867.

Guardian (2010). Greenpeace activists scale bp's london headquarters in oil protest. *Guardian* 20 June.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science* 312(5770), 108–111.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics* 120, 220–230.

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: using natural field experiments to enhance tax compliance. *Journal of Public Economics* 148, 14–31.

Hammerstein, P. (2003). *Genetic and cultural evolution of cooperation*. MIT press.

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics* 4(1), 131–158.

Hanushek, E. A., & Woessmann, L. (2016). Knowledge capital, growth, and the east asian miracle access to schools achieves only so much if quality is poor. *Science* 351(6271), 344–345.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization* 53(1), 3–35.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N. et al. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science* 327(5972), 1480–1484.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N. et al. (2006). Costly punishment across human societies. *Science* 312(5781), 1767–1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science* 319(5868), 1362–1367.

Hobbes, T. (2005 Orig. pub. 1651). *Leviathan*. London: Continuum.

Houser, D., Xiao, E., McCabe, K., & Smith, V. (2008). When punishment fails: research on sanctions, intentions and non-cooperation. *Games and Economic Behavior* 62(2), 509–532.

Isaac, M. R., McCue, K., & Plott, C. R. (1985). Public goods provision in an experimental environment. *Journal of Public Economics* 26, 51–74.

Isaac, M. R., & Walker, J. M. (1984). Divergent evidence on free riding: an experimental examination of some possible explanations. *Public Choice* 43(2), 113–149.

Isaac, R. M., & Walker, J. M. (1988). Communication and free-riding behavior: the voluntary contribution mechanism. *Economic Inquiry* 26(4), 585–608.

Ito, K., Ida, T., & Tanaka, M. (2018). Moral suasion and economic incentives: Field experimental evidence from energy demand. *American Economic Journal: Economic Policy* 10(1), 240–267.

Jensen, K., Call, J., & Tomasello, M. (2007a). Chimpanzees are rational maximizers in an ultimatum game. *Science* 318(5847), 107–109.

Jensen, K., Call, J., & Tomasello, M. (2007b). Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences* 104(32), 13046–13050.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature* 530(7591), 473–476.

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: when norms do and do not affect behavior. *Personality and Social Psychology Bulletin* 26(8), 1002–1012.

Kamei, K. (2017). Altruistic norm enforcement and decision-making format in a dilemma: experimental evidence. *Working Paper*.

Kaplan, R. M., Anderson, J. P., & Kaplan, C. M. (2007). Modeling quality-adjusted life expectancy loss resulting from tobacco use in the united states. *Social Indicators Research* 81(1), 51–64.

Kaur, S. (forthcoming). Nominal wage rigidity in village labor markets. *American Economic Review*.

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science* 322(5908), 1681–1685.

Kessler, G. (2004). *US District Court, D. Columbia, U.S. v. Philip Morris USA Inc. Civil Action No. 99-2496 (GK)*. Report.

Kim, O., & Walker, J. M. (1984). The free rider problem: experimental evidence. *Public Choice* 43(1), 3–24.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association* 14(3), 608–638.

Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. J., & Sutter, M. (2008). Conditional cooperation on three continents. *Economics Letters* 101(3), 175–178.

Kosfeld, M., & Rustagi, D. (2015). Leader punishment and cooperation in groups: experimental field evidence from commons management in ethiopia. *American Economic Review* 105(2), 747–783.

Kotchen, M., & Moon, J. J. (2012). Corporate social responsibility for irresponsibility. *Journal of Economic Analysis & Policy* 12(1), Art. 55.

Krupka, E. L., Leider, S., & Jiang, M. (2016). A meeting of the minds: informal agreements and social norms. *Management Science* 63(6), 1708–1729.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association* 11(3), 495–524.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., & Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences* 111, 6916–6921.

Lewisch, P. G., Ottone, S., & Ponzano, F. (2011). Free-riding on altruistic punishment? an experimental comparison of third-party punishment in a stand-alone and in an in-group environment. *Review of Law & Economics* 7(1), 161–190.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin* 30(9), 1175–1185.

Lindbeck, A., Nyberg, S., & Weibull, J. W. (1999). Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics* 114(1), 1–35.

Lockwood, B. B., Nathanson, C. G., & Weyl, E. G. (2017). Taxation and the allocation of talent. *Journal of Political Economy* 125(5), 1635–1682.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior* 64(1), 237–267.

Lowes, S., Nunn, N., Robinson, J. A., & Weigel, J. L. (2017). The evolution of culture and institutions: evidence from the kuba kingdom. *Econometrica* 85(4), 1065–1091.

Luttmer, E. F., & Singhal, M. (2014). Tax morale. *Journal of Economic Perspectives* 28(4), 149–68.

Makary, M. A., Sexton, J. B., Freischlag, J. A., Holzmueller, C. G., Millman, E. A., Rowen, L., & Pronovost, P. J. (2006). Operating room teamwork among physicians and nurses: teamwork in the eye of the beholder. *Journal of the American College of Surgeons* 202(5), 746–752.

Mankiw, N. G. (2010). Spreading the wealth around: reflections inspired by joe the plumber. *Eastern Economic Journal* 36(3), 285–298.

Markussen, T., Putterman, L., & Tyran, J. R. (2014). Self-organization for collective action: an experimental study of voting on sanction regimes. *Review of Economic Studies* 81(1), 301–324.

Markussen, T., Putterman, L., & Tyran, J.-R. (2016). Judicial error and cooperation. *European Economic Review* 89, 372–388.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J. et al. (2008). More 'altruistic'punishment in larger societies. *Proceedings of the Royal Society of London B: Biological Sciences* 275, 587–592.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences of the United States of America* 108(28), 11375–11380.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* 45(6), 633–644.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal* 38(1), 24–59.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition* 134, 1–10.

Mendes, N., Steinbeis, N., Bueno-Guerra, N., Call, J., & Singer, T. (2018). Preschool children and chimpanzees incur costs to watch punishment of antisocial others. *Nature Human Behaviour* 2(1), 45–51.

Müller, A., & Kräussel, R. (2011). Doing good deeds in times of need: a strategic perspective on corporate disaster donations. *Strategic Management Journal* 32(9), 911–929.

Murphy, K. M., Shleifer, A., & Vishny, R. W. (1991). The allocation of talent: Implications for growth. *Quarterly Journal of Economics* 106(2), 503–530.

Nicklisch, A., Grechenig, K., & Thöni, C. (2015). Information-sensitive leviathans – the emergence of centralized punishment. *WiSo-HH Working Paper Series,(24)*.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics* 92(1), 91–112.

Nikiforakis, N., & Engelmann, D. (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior & Organization* 78(3), 319–332.

Nikiforakis, N., Noussair, C. N., & Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics* 96(9), 797–807.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin* 34(7), 913–923.

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society of London B: Biological Sciences* 276(1655), 323–329.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, 1997. *American Political Science Review* 92(1), 1–22.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives* 14(3), 137–158.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86(2), 404–417.

Philip Morris International Inc. (1999). *PM21 Internal Toolkit*. Report. URL: http://industrydocuments.library.ucsf.edu/tobacco/docs/fgxx0085.

Philip Morris International Inc. (2015). *Form 10-K submitted to US Securities and Exchange Commission*. Report. URL: https://www.pmi.com/investor-relations/reports-filings.

Posner, E. A. (2000). *Law and social norms*. Cambridge, Massachusetts: Harvard University Press.

Proctor, D., Williamson, R. A., Waal, F. B. M., & Brosnan, S. F. (2013). Chimpanzees play the ultimatum game. *Proceedings of the National Academy of Sciences* 110(6), 2070–2075.

Pruckner, G. J., & Sausgruber, R. (2013). Honesty on the streets: A field study on newspaper purchasing. *Journal of the European Economic Association* 11(3), 661–679.

Putterman, L., Tyran, J.-R., & Kamei, K. (2011). Public goods and voting on formal sanction schemes. *Journal of Public Economics* 95(9), 1213–1222.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review* 83(5), 1281–1302.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science* 325(5945), 1272–1275.

Reuben, E., & Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior* 77(1), 122–137.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences* 109(37), 14824–14829.

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444(7120), 718–723.

Rosenblatt, R. (1994). How do they live with themselves? *New York Times Magazine* March 20.

Rothschild, C., & Scheuer, F. (2016). Optimal taxation with rent-seeking. *The Review of Economic Studies* 83(3), 1225–1262.

Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science* 330(6006), 961–965.

Sally, D. (1995). Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Rationality and Society* 7(1), 58–92.

Sefton, M., Shupp, R., & Walker, J. M. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry* 45(4), 671–690.

Sober, E., & Wilson, D. S. (1999). *Unto others: the evolution and psychology of unselfish behavior*. Harvard University Press.

Spar, D. L., & La Mure, L. T. (2003). The power of activism: Assessing the impact of ngos on global business. *California Management Review* 45(3), 78–101.

Steg, L., & Vlek, C. (2009). Encouraging pro-environmental behaviour: An integrative review and research agenda. *Journal of Environmental Psychology* 29(3), 309–317.

Stephens, D. W., McLinn, C. M., & Stevens, J. R. (2002). Discounting and reciprocity in an iterated prisoner's dilemma. *Science* 298(5601), 2216–2218.

Stevens, J. R., & Hauser, M. D. (2004). Why be nice? psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences* 8(2), 60–65.

Sunstein, C. R. (1996). On the expressive function of law. *University of Pennsylvania Law Review* 144(5), 2021–2053.

Sutter, M., Haigner, S., & Kocher, M. G. (2010). Choosing the carrot or the stick? endogenous institutional choice in social dilemma situations. *Review of Economic Studies* 77(4), 1540–1566.

Szczypka, G., Wakefield, M. A., Emery, S., Terry-McElrath, Y. M., Flay, B. R., & Chaloupka, F. J. (2007). Working to make an image: an analysis of three philip morris corporate image media campaigns. *Tobacco Control* 16(5), 344–350.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal* 94(6), 1395–1415.

Traulsen, A., Röhl, T., & Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society of London B: Biological Sciences*.

Tyran, J.-R., & Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *The Scandinavian Journal of Economics* 108(1), 135–156.

Ulber, J., Hamann, K., & Tomasello, M. (2017). Young children, but not chimpanzees, are averse to disadvantageous and advantageous inequities. *Journal of Experimental Child Psychology* 155, 48–66.

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E. et al. (2018). Registered replication report on mazar, amir, and ariely (2008). *Advances in Methods and Practices in Psychological Science*, 2515245918781032.

Wiessner, P. (2005). Norm enforcement among the ju/'hoansi bushmen - a case of strong reciprocity? *Human Nature* 16(2), 115–145.

Wiswall, M., & Zafar, B. (2018). Human capital and expectations about career and family. *Working Paper*.

World Health Organization (2004). *The Tobacco Industry Documents*. Report. URL: http://www.who.int/tobacco/communications/TI_manual_content.pdf.

Worm, B., Hilborn, R., Baum, J. K., Branch, T. A., Collie, J. S., Costello, C., Fogarty, M. J., Fulton, E. A., Hutchings, J. A., Jennings, S. et al. (2009). Rebuilding global fisheries. *Science* 325(5940), 578–585.

Wrong, D. H. (1961). The oversocialized conception of man in modern sociology. *American Sociological Review* 26(2), 183–193.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences USA* 102(20), 7398–7401.

Xiao, E. T. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior* 77(1), 321–344.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51(1), 110–116.

Appendix

Experimental Instructions

1 Instructions Chapter III

Instructions EXO-PRI-DEC

Welcome to Econ-Lab!

Please read the following instructions carefully. If you have any questions, please raise your hand. An assistant will approach you immediately.

General remarks

Today you are taking part in a study at the Department of Economics of the University of Zurich. You will receive a fixed payment of 15 CHF for your participation. Depending on the course of the study, you can earn an additional amount of money. You will receive your payment at the end of the study in cash. Please note that these instructions are exclusively for your private information and that communication is absolutely prohibited during the whole study. If you have any questions, please direct them towards the experimenters. Violating these rules leads to exclusion from this study and all payments. Data collected in this study will at no time be linked to your identity. Your name will be used exclusively for issuing the acknowledgment of your payment. Hence, your anonymity is guaranteed at all times.

Short description of the procedures of the study

At the beginning of the experiment you will be assigned to a group of five participants. Hence, in addition to you, there are four other members in your group. **The group composition does not change over the course of the experiment.** A group consists of **four participants A** and **one participant B**. The assignment to one of these two roles is determined randomly at the beginning of the experiment. Each participant sticks to his/her assigned role until the end of the experiment. The experiment consists of 25 periods. Each period is composed of three phases:

1. In the first phase participants A can decide whether or not to contribute to a common project of the group.
2. In the second phase all group members receive an information about the decisions that the participants have made. **This individual information is accurate with**

a certain probability only. It is therefore possible that the information about the decision of certain group members is false. The specific pieces of information conveyed to participants are independent of each other and can therefore be different from one another.

3. In the third phase each participant A can spend money in order to reduce the income of other participants A of the group. At the end of each period group members are informed about how much their income was reduced in total.

On the next pages we describe the exact procedures of the experiment.

PROCEDURES OF THE STUDY

At the beginning of the experiment you are informed about your randomly assigned role. There are four participants A and one participant B in your group. **You are either a participant A or the participant B. The assignment to one of those two roles is fixed for the whole duration of the experiment.**

You will receive experimental currency units (so-called Token) over the course of the experiment. **1 Token corresponds to 0.05 CHF**. The experiment consists of 25 periods. **At the end of the experiment, the sum of all the Token you have collected over the 25 periods is converted to CHF and paid out to you. You receive this amount in addition to your fixed payment.** The income of a single period may be negative. **You receive an additional payment of 0 CHF if the sum of all your period incomes is negative (your fixed payment of 15 CHF remains unaffected).**

Participant B

As participant B you do not make any decisions during the experiment. However, in each period you receive a share of the project's profit, which depends on the decisions of participants A. The profit of the project is split equally among all five group members. In each period you receive information about whether the participants A contributed to the project. Details can be found below in the section "phases of the experiment".

Participant A

As one of the four participants A you decide in each of the 25 periods whether or not you want to contribute to the common project. In each period you receive a share of the project's profit that depends on the decisions of all participants A. The profit of the project is split equally among all five group members. Furthermore, like participant B, you receive information about whether or not the other participants A contributed to the project. Afterwards, you can punish particular participants A by reducing their income. Details can be found below in the section "phases of the experiment".

Phases of the experiment

Phase 1 – Decision about the contribution to the project

As participant A you are endowed with an amount of 15 Token at the beginning of each period. You have to choose one of two options:

1. Either you contribute the 15 Token to the common project or
2. You keep the 15 Token.

All group members (participants A as well as participant B) profit equally from contributions to the common project. The sum of contributions is first doubled and then split equally among all group members.

If a group member contributes the 15 Token to the project, the income from the project increases by $\frac{(15*2)}{5} = 6$ Token for each group member.

As participant B you receive a share of the project's income, but you do not make any decisions yourself.

In this phase, you are not informed about the actual income from the project - neither as participant A nor as participant B. Only at the end of the experiment (i.e. after 25 periods) all participants are informed about their total earnings in each period.

project income per group member = number of contributing participants x 6 Token

Examples:

- Suppose all participants A contribute 15 Token to the common project. Then each group member gets 6 Token * 4 = 24 Token from the project.
- Suppose no one contributes to the project. Then all participants A keep their 15 Token and participant B receives 0 additional Token.
- Suppose you are participant A and you do not contribute to the project. Each of the other three participants A contributes 15 Token to the project. Then you keep your 15 Token and you receive an additional 3*6 Token = 18 Token from the project. Hence, in total you get 33 Token. Each other group member receives 3*6 Token = 18 Token from the project.

Phase 2 – Information about contributions of other participants:

In this phase all group members (participants A and B) individually receive an independent information about each participant A's contribution decision. **Each piece of information is correct with a probability of 90% and false with a probability of 10%. Hence,**

- when a participant A actually contributes 15 Token, then you receive the information "15 Token" with 90% probability and the information "0 Token" with 10% probability.
- when a participant A actually contributes 0 Token, then you receive the information "0 Token" with 90% probability and the information "15 Token" with 10% probability.

The pieces of information are provided individually and are independent of each other. **Therefore, it is possible that two group members receive different information about the actions of one and the same participant A.** You do not learn the information received by other group members.

The labels of participants A are randomly reassigned each period. For example, "participant A2" in the first period is not necessarily the same person as "participant A2" in the second period.

Phase 3 – Decision about punishment of other participants:

In this phase participants A have the possibility to punish other participants A by reducing their period income.

If you as a participant A decide to punish another participant A, then 8 Token are deducted from his/her period income. You have to pay 2 Token in order to punish another participant A. For that purpose you are endowed with an additional 6 Token in this phase.

Hence, you can punish a) no participant A, b) one participant A, c) two participants A or d) all three participants A.

Participant B has to bear costs of 2 Token for each exerted punishment as well. For that purpose participant B is endowed with additional 24 Token. However, participant B does not make any punishment decisions. At the end of each period all participants A are informed about the punishment they have received. They learn by how many participants they were punished and by how much their income was reduced in total.

Overview of total income within one period

period income of a participant A

=

(15 Token) – (contribution to project)

+

(6 Token x number of contributing participants)

+

(6 Token) – (2 Token x number of assigned punishments)

-

(8 Token x number of received punishments)

period income of participant B

=

(6 Token x number of contributing participants)

+

(24 Token) – (2 Token x sum of all assigned punishments)

Control questions

Please answer the following questions and raise your hand.

1. Question

Suppose you are participant A and no participant A has contributed to the common project. Suppose further that no participant A makes use of the possibility to punish. What is you period income (please keep in mind that you are endowed with an additional 6 Token in phase 3)

Your period income (in Token) _____

2. Question

Suppose you are a participant A and you have contributed to the common project. In addition to you, two other participants A have contributed to the project. Suppose further that you punish one participant A and that you do not receive any punishment yourself.

a) What is your period income? Your period income (in Token) _____

b) Now suppose that you are punished by two participants A. What is your period income in this case?

Your period income (in Token) _____

3. Question

Suppose you receive the information that participant A2 has not contributed to the project.

a) What is the probability that this information about participant A2 is correct?

_____

b) Do all group members necessarily receive the same information about the contribution of participant A2?

_____YES _____NO

c) Next period you receive new information about the contribution of participant A2. Does participant A2 necessarily correspond the same actual person as in the previous period?

_____YES _____NO

Instructions EXO-PRI-CEN

Welcome to Econ-Lab!

Please read the following instructions carefully. If you have any questions, please raise your hand. An assistant will approach you immediately.

General remarks

Today you are taking part in a study at the Department of Economics of the University of Zurich. You will receive a fixed payment of 15 CHF for your participation. Depending on the course of the study, you can earn an additional amount of money. You will receive your payment at the end of the study in cash. Please note that these instructions are exclusively for your private information and that communication is absolutely prohibited during the whole study. If you have any questions, please direct them towards the experimenters. Violating these rules leads to exclusion from this study and all payments. Data collected in this study will at no time be linked to your identity. Your name will be used exclusively for issuing the acknowledgment of your payment. Hence, your anonymity is guaranteed at all times.

Short description of the procedures of the study

At the beginning of the experiment you will be assigned to a group of five participants. Hence, in addition to you, there are four other members in your group. **The group composition does not change over the course of the experiment.** A group consists of **four participants A** and **one participant B**. The assignment to one of these two roles is determined randomly at the beginning of the experiment. Each participant sticks to his/her assigned role until the end of the experiment. The experiment consists of 25 periods. Each period is composed of three phases:

1. In the first phase participants A can decide whether or not to contribute to a common project of the group.
2. In the second phase all group members receive an information about the decisions that the participants have made. **This individual information is accurate with a certain probability only.** It is therefore possible that the information about the decision of certain group members is false. The specific pieces of information

conveyed to participants are independent of each other and can therefore be different from one another.

3. In the third phase each participant B can spend money in order to reduce the income of participants A of the group. At the end of each period group members are informed about how much their income was reduced in total.

On the next pages we describe the exact procedures of the experiment.

PROCEDURES OF THE STUDY

At the beginning of the experiment you are informed about your randomly assigned role. There are four participants A and one participant B in your group. **You are either a participant A or the participant B. The assignment to one of those two roles is fixed for the whole duration of the experiment.**

You will receive experimental currency units (so-called Token) over the course of the experiment. **1 Token corresponds to 0.05 CHF**. The experiment consists of 25 periods. **At the end of the experiment, the sum of all the Token you have collected over the 25 periods is converted to CHF and paid out to you. You receive this amount in addition to your fixed payment.** The income of a single period may be negative. **You receive an additional payment of 0 CHF if the sum of all your period incomes is negative (your fixed payment of 15 CHF remains unaffected).**

Participant B

As participant B you do not make any decisions during the first two phases of the experiment. However, in each period you receive a share of the project's profit, which depends on the decisions of participants A. The profit of the project is split equally among all five group members. In each period you receive information about whether the participants A contributed to the project. In the third phase of each period you can punish participants A by reducing their income. Details can be found below in the section "phases of the experiment".

Participant A

As one of the four participants A you decide in each of the 25 periods whether or not you want to contribute to the common project. In each period you receive a share of the project's profit that depends on the decisions of all participants A. The profit of the project is split equally among all five group members. Furthermore, like participant B, you receive information about whether or not the other participants A contributed to the project. In the third phase you do not make any decision. You are, however, informed whether and how severe you were punished by participant B. Details can be found below in the section "phases of the experiment".

Phases of the experiment

Phase 1 – Decision about the contribution to the project

As participant A you are endowed with an amount of 15 Token at the beginning of each period. You have to choose one of two options:

1. Either you contribute the 15 Token to the common project or
2. You keep the 15 Token.

All group members (participants A as well as participant B) profit equally from contributions to the common project. The sum of contributions is first doubled and then split equally among all group members.

If a group member contributes the 15 Token to the project, the income from the project increases by $\frac{(15*2)}{5} = 6$ Token for each group member.

As participant B you receive a share of the project's income, but you do not make any decisions yourself.

In this phase, you are not informed about the actual income from the project - neither as participant A nor as participant B. Only at the end of the experiment (i.e. after 25 periods) all participants are informed about their total earnings in each period.

project income per group member = number of contributing participants x 6 Token

Examples:

- Suppose all participants A contribute 15 Token to the common project. Then each group member gets 6 Token * 4 = 24 Token from the project.
- Suppose no one contributes to the project. Then all participants A keep their 15 Token and participant B receives 0 additional Token.
- Suppose you are participant A and you do not contribute to the project. Each of the other three participants A contributes 15 Token to the project. Then you keep your 15 Token and you receive an additional 3*6 Token = 18 Token from the project. Hence, in total you get 33 Token. Each other group member receives 3*6 Token = 18 Token from the project.

Phase 2 – Information about contributions of other participants:

In this phase all group members (participants A and B) individually receive an independent information about each participant A's contribution decision. **Each piece of information is correct with a probability of 90% and false with a probability of 10%. Hence,**

- when a participant A actually contributes 15 Token, then you receive the information "15 Token" with 90% probability and the information "0 Token" with 10% probability.
- when a participant A actually contributes 0 Token, then you receive the information "0 Token" with 90% probability and the information "15 Token" with 10% probability.

The pieces of information are provided individually and are independent of each other. **Therefore, it is possible that two group members receive different information about the actions of one and the same participant A.** You do not learn the information received by other group members.

The labels of participants A are randomly reassigned each period. For example, "participant A2" in the first period is not necessarily the same person as "participant A2" in the second period.

Phase 3 – Decision about punishment of other participants:

In this phase participant B has the possibility to punish participants A by reducing their period income.

As a participant B you decide in every period whether to reduce a participant A's income by a) 0 Token, b) 8 Token, c) 16 Token or d) 24 Token. You have to pay 1 Token per every 4 Token that are deducted from a participant A's income. For that purpose you are endowed with additional 24 Token in this phase.

The other three participants A have to bear total costs of 1 Token per 4 Token that are deducted from the fourth participant A. These total costs are split equally among the three other participants A. For that purpose each participant A is endowed with an additional 6 Token.

If for example, participant B reduces the income of one participant A by 24 Token then the other three participants A have to pay 6 Token in total, i.e. 2 Token each.

At the end of each period all participants A are informed about the punishment they have received. They do not learn about the punishment of other participants A.

Overview of total income within one period

period income of a participant A

=

(15 Token) – (contribution to project)

+

(6 Token x number of contributing participants)

+

(6 Token)

-

($\frac{1}{4}$x$\frac{1}{3}$ Token x sum of punishment of other participants A by B)

-

(received punishments by participant B)

period income of participant B

=

(6 Token x number of contributing participants)

+

(24 Token)

–

($\frac{1}{4}$ Token x sum of all assigned punishments to A)

Control questions

Please answer the following questions and raise your hand.

1. Question

Suppose you are participant A and no participant A has contributed to the common project. Suppose participant B does not make use of the possibility to punish.

What is you period income (please keep in mind that you are endowed with an additional 6 Token in phase 3)

Your period income (in Token) _____

2. Question

Suppose you are a participant A and you have contributed to the common project. The other three participants A have also contributed to the project. Suppose further that you are punished with 24 Token (your income is reduced by 24 Token) by participant B and that no other participant A is punished.

What is your period income? Your period income (in Token) _____

3. Question

Suppose you receive the information that participant A2 has not contributed to the project.

a) What is the probability that this information about participant A2 is correct?

_____

b) Do all group members necessarily receive the same information about the contribution of participant A2?

_____YES _____NO

c) Next period you receive new information about the contribution of participant A2. Does participant A2 necessarily correspond the same actual person as in the previous period?

_____YES _____NO

4. Question Suppose you are participant B and two participants A have contributed to the common project. Suppose further that you punish two participants with 24 Token (you reduce the income of two participants A by 24 Token).

What is your period income? Your period income (in Token)_____

Instructions END-PRI-DEC

Welcome to Econ-Lab!

Please read the following instructions carefully. If you have any questions, please raise your hand. An assistant will approach you immediately.

General remarks

Today you are taking part in a study at the Department of Economics of the University of Zurich. You will receive a fixed payment of 15 CHF for your participation. Depending on the course of the study, you can earn an additional amount of money. You will receive your payment at the end of the study in cash. Please note that these instructions are exclusively for your private information and that communication is absolutely prohibited during the whole study. If you have any questions, please direct them towards the experimenters. Violating these rules leads to exclusion from this study and all payments. Data collected in this study will at no time be linked to your identity. Your name will be used exclusively for issuing the acknowledgment of your payment. Hence, your anonymity is guaranteed at all times.

Short description of the procedures of the study

At the beginning of the experiment you will be assigned to a group of five participants. Hence, in addition to you, there are four other members in your group. **The group composition does not change over the course of the experiment.** A group consists of **four participants A** and **one participant B**. The assignment to one of these two roles is determined randomly at the beginning of the experiment. Each participant sticks to his/her assigned role until the end of the experiment. The experiment consists of 25 periods. Each period is composed of three phases:

1. In the first phase participants A can decide whether or not to contribute to a common project of the group.
2. In the second phase all group members receive an information (free of charge) about the decisions that the participants have made. **This individual information is accurate with a certain probability only.** It is therefore possible that the information about the decision of certain group members is false. The specific

pieces of information conveyed to participants are independent of each other and can therefore be different from one another. In order to improve your information base you can buy further information about other group members. The new information is also accurate with a certain probability only.

3. In the third phase each participant A can spend money in order to reduce the income of other participants A of the group. At the end of each period group members are informed about how much their income was reduced in total.

On the next pages we describe the exact procedures of the experiment.

PROCEDURES OF THE STUDY

At the beginning of the experiment you are informed about your randomly assigned role. There are four participants A and one participant B in your group. **You are either a participant A or the participant B. The assignment to one of those two roles is fixed for the whole duration of the experiment.**

You will receive experimental currency units (so-called Token) over the course of the experiment. **1 Token corresponds to 0.05 CHF**. The experiment consists of 25 periods. **At the end of the experiment, the sum of all the Token you have collected over the 25 periods is converted to CHF and paid out to you. You receive this amount in addition to your fixed payment.** The income of a single period may be negative. **You receive an additional payment of 0 CHF if the sum of all your period incomes is negative (your fixed payment of 15 CHF remains unaffected).**

Participant B

As participant B you do not make any decisions during the experiment. However, in each period you receive a share of the project's profit, which depends on the decisions of participants A. The profit of the project is split equally among all five group members. In each period you receive information about whether the participants A contributed to the project. Details can be found below in the section "phases of the experiment".

Participant A

As one of the four participants A you decide in each of the 25 periods whether or not you want to contribute to the common project. In each period you receive a share of the project's profit that depends on the decisions of all participants A. The profit of the project is split equally among all five group members. Furthermore, like participant B, you receive information about whether or not the other participants A contributed to the project. Afterwards, you can acquire further information about the behavior of other participants A in order to improve your information base. Finally, you can punish particular participants A by reducing their income. Details can be found below in the section "phases of the experiment".

Phases of the experiment

Phase 1 – Decision about the contribution to the project

As participant A you are endowed with an amount of 15 Token at the beginning of each period. You have to choose one of two options:

1. Either you contribute the 15 Token to the common project or
2. You keep the 15 Token.

All group members (participants A as well as participant B) profit equally from contributions to the common project. The sum of contributions is first doubled and then split equally among all group members.

If a group member contributes the 15 Token to the project, the income from the project increases by $\frac{(15*2)}{5} = 6$ Token for each group member.

As participant B you receive a share of the project's income, but you do not make any decisions yourself.

In this phase, you are not informed about the actual income from the project - neither as participant A nor as participant B. Only at the end of the experiment (i.e. after 25 periods) all participants are informed about their total earnings in each period.

project income per group member = number of contributing participants x 6 Token

Examples:

- Suppose all participants A contribute 15 Token to the common project. Then each group member gets 6 Token * 4 = 24 Token from the project.
- Suppose no one contributes to the project. Then all participants A keep their 15 Token and participant B receives 0 additional Token.
- Suppose you are participant A and you do not contribute to the project. Each of the other three participants A contributes 15 Token to the project. Then you keep your 15 Token and you receive an additional 3*6 Token = 18 Token from the project. Hence, in total you get 33 Token. Each other group member receives 3*6 Token = 18 Token from the project.

Phase 2 – Information about contributions of other participants:

In this phase all group members (participants A and B) individually receive an independent information about each participant A's contribution decision. **Each piece of information is correct with a probability of 90% and false with a probability of 10%. Hence,**

- when a participant A actually contributes 15 Token, then you receive the information "15 Token" with 90% probability and the information "0 Token" with 10% probability.
- when a participant A actually contributes 0 Token, then you receive the information "0 Token" with 90% probability and the information "15 Token" with 10% probability.

The pieces of information are provided individually and are independent of each other. **Therefore, it is possible that two group members receive different information about the actions of one and the same participant A.** You do not learn the information received by other group members.

Afterwards, all four participants A have the possibility to acquire up to two further pieces of information about each of the other three participants A. Hence, in total each participant A can buy 6 additional pieces of information (2 pieces of information x 3 other participants A). One new piece of information costs 1 Token for the acquiring participant A. Participant B also has to bear costs of 1 Token for each new piece of information that is acquired.

Like the initial three pieces of costless information, all new pieces of information are independent of each other and with 90% probability true and with 10% probability false. New pieces of information are acquired sequentially, i.e. one after another. You can stop to buy further information at any time.

The labels of participants A are randomly reassigned each period. For example, "participant A2" in the first period is not necessarily the same person as "participant A2" in the second period.

Phase 3 – Decision about punishment of other participants:

In this phase participants A have the possibility to punish other participants A by reducing their period income.

If you as a participant A decide to punish another participant A, then 8 Token are deducted from his/her period income. You have to pay 2 Token in order to punish another participant A. Hence, you can punish a) no participant A, b) one participant A, c) two participants A or d) all three participants A.

Participant B has to bear costs of 2 Token for each exerted punishment as well. However, participant B does not make any punishment decisions.

At the end of each period all participants A are informed about the punishment they have received. They learn by how many participants they were punished and by how much their income was reduced in total.

Overview of total income within one period

period income of a participant A

=

(15 Token) - (contribution to project)

+

(6 Token x number of contributing participants)

+

(6 Token) - (1 Token x number of acquired pieces of information)

-

(2 Token x number of assigned punishments)

-

(8 Token x number of received punishments)

period income of participant B

=

(6 Token x number of contributing participants)

+

(24 Token)

-

(1 Token x total number of acquired pieces of information)

–

(2 Token x sum of all assigned punishments)

Control questions

Please answer the following questions and raise your hand.

1. Question

Suppose you are participant A and no participant A has contributed to the common project. Suppose further that no participant A makes use of the possibility to punish. What is you period income if no further information is acquired (please keep in mind

that you are endowed with an additional 6 Token in phase 3)?
Your period income (in Token) _____

2. Question
Suppose you are a participant A and you have contributed to the common project. In addition to you, two other participants A have contributed to the project. Suppose further that you punish one participant A and that you do not receive any punishment yourself. Furthermore, you buy one additional piece of information for each of the other three participants A.
a) What is your period income? Your period income (in Token) _____
b) Now suppose that you are punished by two participants A. What is your period income in this case?
Your period income (in Token) _____

3. Question
Suppose you receive the information that participant A2 has not contributed to the project.
a) What is the probability that this information about participant A2 is correct?

_____

b) Do all group members necessarily receive the same information about the contribution of participant A2?
_____YES _____NO
c) Suppose you want to acquire one additional piece of information about participant A2. How much do you have to pay for that? _____Token What is the probability that this newly acquired information about participant A2 is correct?

_____

d) Next period you receive new information about the contribution of participant A2. Does participant A2 necessarily correspond the same actual person as in the previous period?
_____YES _____NO

Instructions END-PRI-CEN

Welcome to Econ-Lab!

Please read the following instructions carefully. If you have any questions, please raise your hand. An assistant will approach you immediately.

General remarks

Today you are taking part in a study at the Department of Economics of the University of Zurich. You will receive a fixed payment of 15 CHF for your participation. Depending on the course of the study, you can earn an additional amount of money. You will receive your payment at the end of the study in cash. Please note that these instructions are exclusively for your private information and that communication is absolutely prohibited during the whole study. If you have any questions, please direct them towards the experimenters. Violating these rules leads to exclusion from this study and all payments. Data collected in this study will at no time be linked to your identity. Your name will be used exclusively for issuing the acknowledgment of your payment. Hence, your anonymity is guaranteed at all times.

Short description of the procedures of the study

At the beginning of the experiment you will be assigned to a group of five participants. Hence, in addition to you, there are four other members in your group. **The group composition does not change over the course of the experiment.** A group consists of **four participants A** and **one participant B**. The assignment to one of these two roles is determined randomly at the beginning of the experiment. Each participant sticks to his/her assigned role until the end of the experiment. The experiment consists of 25 periods. Each period is composed of three phases:

1. In the first phase participants A can decide whether or not to contribute to a common project of the group.
2. In the second phase all group members receive an information (free of charge) about the decisions that the participants have made. **This individual information is accurate with a certain probability only.** It is therefore possible that the information about the decision of certain group members is false. The specific

pieces of information conveyed to participants are independent of each other and can therefore be different from one another. In order to improve your information base participant B can buy further information about other group members. The new information is also accurate with a certain probability only.

3. In the third phase each participant B can spend money in order to reduce the income of participants A of the group. At the end of each period group members are informed about how much their income was reduced in total.

On the next pages we describe the exact procedures of the experiment.

PROCEDURES OF THE STUDY

At the beginning of the experiment you are informed about your randomly assigned role. There are four participants A and one participant B in your group. **You are either a participant A or the participant B. The assignment to one of those two roles is fixed for the whole duration of the experiment.**

You will receive experimental currency units (so-called Token) over the course of the experiment. **1 Token corresponds to 0.05 CHF**. The experiment consists of 25 periods. **At the end of the experiment, the sum of all the Token you have collected over the 25 periods is converted to CHF and paid out to you. You receive this amount in addition to your fixed payment.** The income of a single period may be negative. **You receive an additional payment of 0 CHF if the sum of all your period incomes is negative (your fixed payment of 15 CHF remains unaffected).**

Participant B

As participant B you do not make any decisions during the first phase of the experiment. However, in each period you receive a share of the project's profit, which depends on the decisions of participants A. The profit of the project is split equally among all five group members. In each period you receive information about whether the participants A contributed to the project. Afterwards, you can acquire further information about the behavior of other participants A in order to improve your information base. In the third phase of each period you can punish participants A by reducing their income. Details can be found below in the section "phases of the experiment".

Participant A

As one of the four participants A you decide in each of the 25 periods whether or not you want to contribute to the common project. In each period you receive a share of the project's profit that depends on the decisions of all participants A. The profit of the project is split equally among all five group members. Furthermore, like participant B, you receive information about whether or not the other participants A contributed to the project. In the third phase you do not make any decision. Details can be found below in the section "phases of the experiment".

Phases of the experiment

Phase 1 – Decision about the contribution to the project

As participant A you are endowed with an amount of 15 Token at the beginning of each period. You have to choose one of two options:

1. Either you contribute the 15 Token to the common project or
2. You keep the 15 Token.

All group members (participants A as well as participant B) profit equally from contributions to the common project. The sum of contributions is first doubled and then split equally among all group members.

If a group member contributes the 15 Token to the project, the income from the project increases by $\frac{(15*2)}{5} = 6$ Token for each group member.

As participant B you receive a share of the project's income, but you do not make any decisions yourself.

In this phase, you are not informed about the actual income from the project - neither as participant A nor as participant B. Only at the end of the experiment (i.e. after 25 periods) all participants are informed about their total earnings in each period.

project income per group member = number of contributing participants x 6 Token

Examples:

- Suppose all participants A contribute 15 Token to the common project. Then each group member gets 6 Token * 4 = 24 Token from the project.
- Suppose no one contributes to the project. Then all participants A keep their 15 Token and participant B receives 0 additional Token.
- Suppose you are participant A and you do not contribute to the project. Each of the other three participants A contributes 15 Token to the project. Then you keep your 15 Token and you receive an additional 3*6 Token = 18 Token from the project. Hence, in total you get 33 Token. Each other group member receives 3*6 Token = 18 Token from the project.

Phase 2 – Information about contributions of other participants:

In this phase all group members (participants A and B) individually receive an independent information about each participant A's contribution decision. **Each piece of information is correct with a probability of 90% and false with a probability of 10%. Hence,**

- when a participant A actually contributes 15 Token, then you receive the information "15 Token" with 90% probability and the information "0 Token" with 10% probability.
- when a participant A actually contributes 0 Token, then you receive the information "0 Token" with 90% probability and the information "15 Token" with 10% probability.

The pieces of information are provided individually and are independent of each other. Therefore, it is possible that two group members receive different information about the actions of one and the same participant A. You do not learn the information received by other group members.

Afterwards, participant B has the possibility to acquire up to two further pieces of information about each of the four participants A. Hence, in total each participant A can buy 8 additional pieces of information (2 pieces of information x 3 other participants A). One new piece of information costs 3 Token for participant B.

Information acquired by participant B is conveyed to the other three participants A. Each participant A also has to bear costs of 1 Token for each new piece of information that he or she receives.

Like the initial three pieces of costless information, all new pieces of information are independent of each other and with 90% probability true and with 10% probability false. New pieces of information are acquired sequentially, i.e. one after another. You can stop to buy further information at any time.

The labels of participants A are randomly reassigned each period. For example, "participant A2" in the first period is not necessarily the same person as "participant A2" in the second period.

Phase 3 – Decision about punishment of other participants:

In this phase participant B has the possibility to punish participants A by reducing their period income.

As a participant B you decide in every period whether to reduce a participant A's income by a) 0 Token, b) 8 Token, c) 16 Token or d) 24 Token. You have to pay 1 Token per every 4 Token that are deducted from a participant A's income. For that purpose you are endowed with additional 24 Token in this phase.

The other three participants A have to bear **total** costs of 1 Token per 4 Token that are deducted from the fourth participant A. These total costs are split equally among the three other participants A. For that purpose each participant A is endowed with an additional 6 Token.

If for example, participant B reduces the income of one participant A by 24 Token then the other three participants A have to pay 6 Token in total, i.e. 2 Token each.

At the end of each period all participants A are informed about the punishment they have received. They do not learn the punishment of other participants A.

Overview of total income within one period

period income of a participant A

=

(15 Token) – (contribution to project) +

(6 Token x number of contributing participants)

+

(6 Token)

–

(1 Token x further pieces of information about participants A)

-

($\frac{1}{12}$ Token x sum of total punishment of other participants A by B)

-

(received punishment by participant B)

period income of participant B

=

(6 Token x number of contributing participants)

+

(24 Token)

-

(3 Token x number of acquired pieces of information)

–

($\frac{1}{4}$ Token x sum of all assigned punishment to participants A)

Control questions

Please answer the following questions and raise your hand.

1. Question

Suppose you are participant A and no participant A has contributed to the common project. Suppose further that no participant A makes use of the possibility to punish.

What is you period income if participant B does not acquire further information (please keep in mind that you are endowed with an additional 6 Token in phase 3)?
Your period income (in Token) _____

2. Question
Suppose you are a participant A and you have contributed to the common project. In addition to you, the other three participants A have also contributed to the project. Suppose further that you are punished by participant B with 24 Token (your period income is reduced by 24 Token). Furthermore no other participant A is punished.
What is your period income? Your period income (in Token) _____

3. Question
Suppose you receive the information that participant A2 has not contributed to the project.
a) What is the probability that this information about participant A2 is correct? _____
b) Do all group members necessarily receive the same information about the contribution of participant A2?
_____YES _____NO
c) Next period you receive new information about the contribution of participant A2. Does participant A2 necessarily correspond the same actual person as in the previous period?
_____YES _____NO

4. Question Suppose you are participant B and two participants A have contributed to the common project. Suppose further that you punish two participants with 24 Token (you reduce the income of two participants A by 24 Token). Additionally you acquire one further piece of information for each participant A.
a) What is your period income? Your period income (in Token)_____
b) What is the probability that the newly acquired piece of information about participant A2 is correct? _____

2 Instructions Chapter IV

Instructions sessions NF–noNF Part 1

Welcome

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

General information

You are now participating in an economic experiment. You will receive a fixed amount of 5 Pounds Sterling. During the study you will be able to earn more money. You will receive your earnings in cash at the end of the study.

During the experiment we talk about Token instead of Pounds Sterling. Initially, your earnings will therefore be calculated in Token. At the end of the experiment, the total sum of Token is converted into Pounds. The following condition will hold:

$$1 \text{ Token} = 1\text{p}$$

Every participant will get (additional to the show-up fee of £5) a one-time lump sum payment of **200 Token**. With this lump sum payment you will be able to cover possible losses. At the end of the experiment you will receive your total sum of Token (including the lump sum payment) in addition to the £4 show-up fee. Your earnings will be paid out in cash.

You are not allowed to communicate with the other participants. Please ask the experimenter if you have any questions. The violation of this rule will lead to the exclusion of the experiment and of all the above mentioned payments.

The data collected during the study will not be matched with your identity at any point.

Short description of the study

At the beginning of the experiment you will be randomly assigned to a group of four. Hence, there will be three other participants in the group with you. **The group com-**

position will not change during the course of the study. You will only interact with the members of your own group. Every group member has the same possibility as the other members and will receive the same instructions.

The experiment consists of two parts. The instructions for the second part will be handed out after the conclusion of the first part. Your total income will be a sum of the two parts. The first part of the experiment consists of 15 periods. Each of the 4 group members will have to decide in each period how many Token they want to contribute to the project. Each group member can contribute between 0 and 20 Token to the project. Each period consists of 4 stages:

1. During the first stage each group member has the opportunity to express how much he or she thinks that every group member should contribute to the project.
2. During the second stage each group member decides how many Token he or she will contribute to the project.
3. During the third stage every group member will be informed about how many Token will have been contributed by the other group members. Afterwards the members will be able to spend Token in order to reduce the earnings of the other group members.
4. During the final stage the group members will again get the chance to spend Token in order to reduce the earnings of the other group members. They will, however, only be able to reduce the earnings of those group members, who reduced their earnings during the third stage.

At the end of each period you will be informed about how much you will have earned during this period and about the composition of these earnings. On the following pages, we will describe the exact procedure of the experiment.

Procedure of the study

At the beginning of the first part you will be randomly assigned to a group of four. Hence, you and three other participants will together form one group. These groups will remain unchanged for the whole experiment.

At the beginning of each period – during stage 1 – you will be able to indicate how much each group member should contribute to the project in your opinion. The average value

will be conveyed to all members of your group. During stage 2 you will decide how much you want to contribute to the project. You will receive a share of the earnings of the project, which in turn depends on the decisions of all group members. The earnings of the project will be divided equally among all four group members. Further details will be described below. During stage 3 you will be informed about how much the other group members will indeed have contributed to the project. In addition, you will be able to use your Token in order to reduce the earnings of the other group members during this stage. This will be possible through the assignment of reduction points. During stage 4 the members whose earnings were reduced by other group members during stage 3, will in turn be able to assign counter reduction points to those and only those group members. During the first part of the experiment these 4 stages will be repeated 15 times. Afterwards the instructions for the second part will be distributed.

The stages of the experiment in detail

Stage 1 – Communication about how much each member should contribute to the project

At the beginning of each period you will be asked the following questions (see below the monitor screen for stage 1):

"In your opinion, how many Token should each group member contribute to the project?"

Since every group member can contribute between 0 and 20 Token to the project, you have to answer this question with a whole number from the range of 0 to 20.

Screen stage 1: Input of your answer into the empty box and confirmation with "Continue"

The average of the answers of your group will be calculated and subsequently conveyed to each group member. The average will be rounded to the nearest whole number. This information will be available on the monitor screen of stage 2 (see below).

Stage 2 – Decision about how much to contribute to the project

In every period each group member will get 20 Token. Each group member has to decide how much he or she wants to contribute to the project. Every whole number between 0 and 20 can be chosen. Each group member profits equally from the earnings of the project. **The sum of the contributed Token will be multiplied by 1.6 (+60% of all the contributions) and will be equally redistributed**.

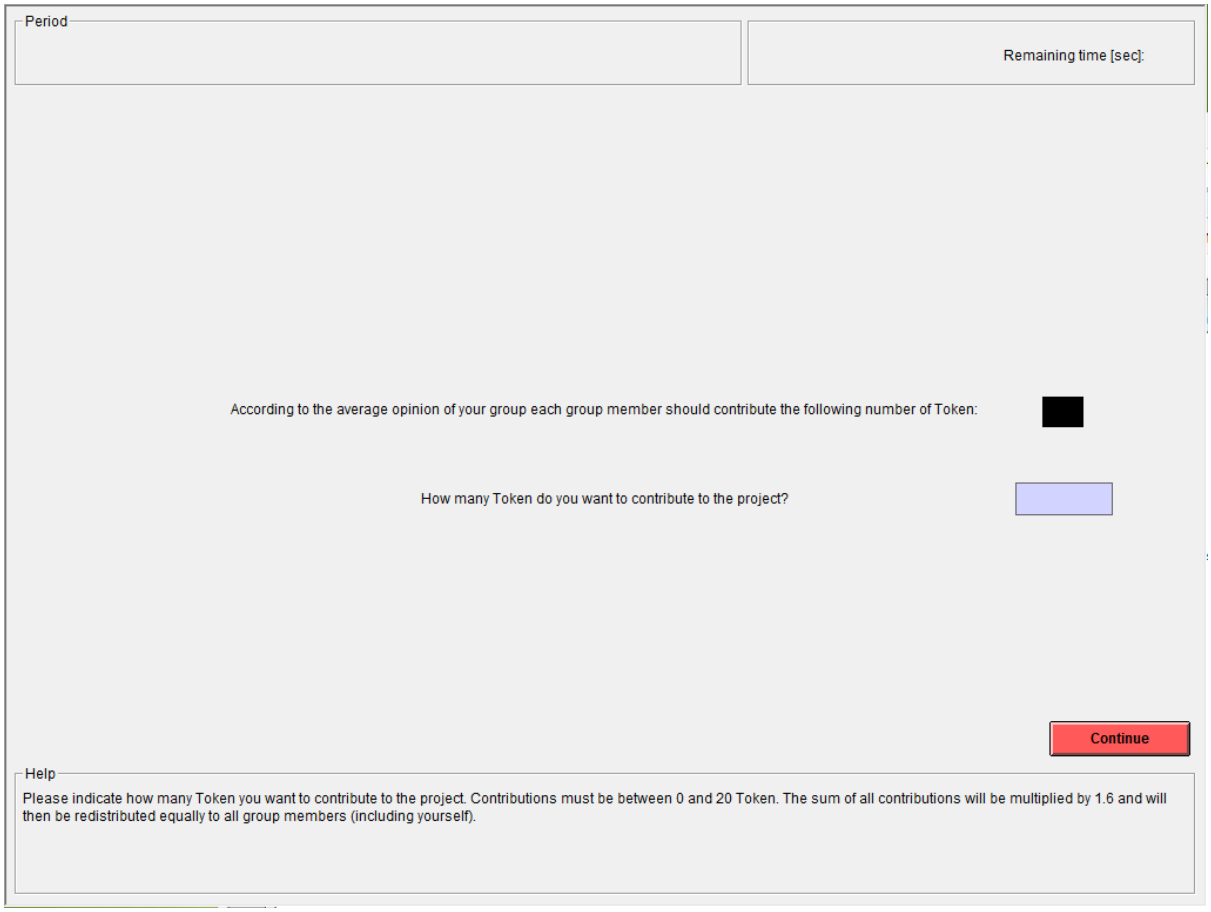
The earnings of the project can be calculated by 1.6*X Token, X Token being the sum of all contributions. This amount will be equally redistributed to all group members. In other words, each group member will receive $\frac{1.6*X}{4} = 0.4$ X Token. Therefore, for every

contributed Token, each group member will receive 0.4 Token (including you). You can keep the Token, which you will not have contributed, for yourself. These Token will be part of your total earnings. The input is made as shown in the monitor screen below.

Earnings from the project for each group member = 0.4 * amount of contributed Tokens

Examples:

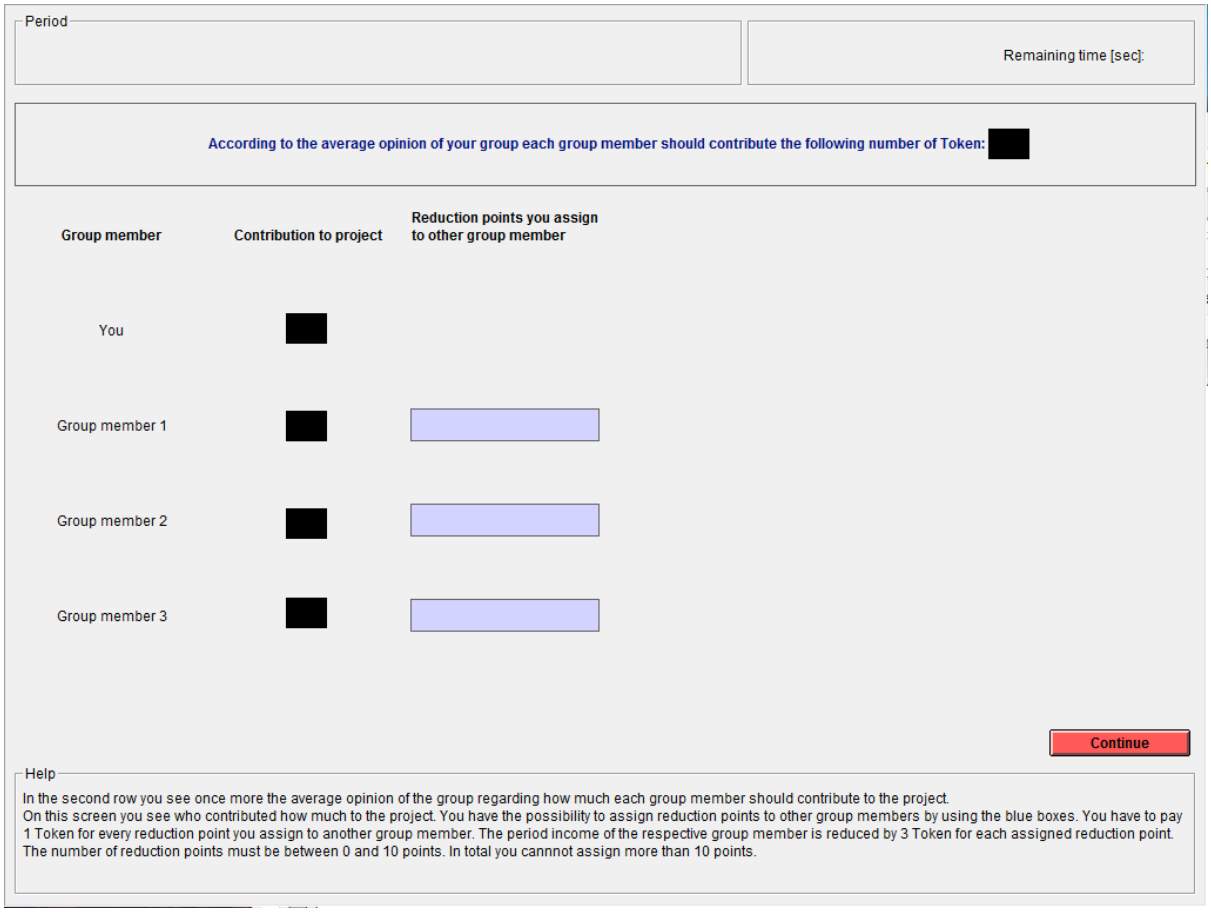
- Assuming each group member will contribute 20 Token to the project, 80 Token will be available for the project in total. Each group member will receive 0.4*80 = 32 Token from the project.
- Assuming nobody will contribute to the project (0 Token), then nobody will receive earnings from the project since every group member decided to keep 20 Token for him- or herself.
- Assuming you contribute 5 Token to the project and each of the other group members contributes 10 Token, then you will get 0.4*(5+10+10+10) = 14 Token in addition to the 15 Token (which you did not contribute). The other group members will get 14 Token from the project as well, but they will only have 10 Token left from before.



Screen stage 2: Input of your answer into the empty box and confirmation with "Continue"

Stage 3 – Assignment of reduction points

At the beginning of this stage each group member will receive 10 additional Token. These Token can be used to reduce the earnings of the other members of the group. You can do this by assigning reduction points. At the beginning of this stage you will see how much the other group members will have contributed to the project (see monitor screen below). Afterwards, you will decide whether or not you want to assign reduction points to other group members, and in case you want to do so, how many reduction points you want to assign. You have to pay 1 Token for each reduction point you want to assign. The group member's earnings will be reduced by 3 Token for every reduction point received. More specifically, you can pay 1 Token to reduce the earnings of another member by 3 Tokens. You can distribute a maximum of 10 reduction points. The other members have the same possibility as you do.



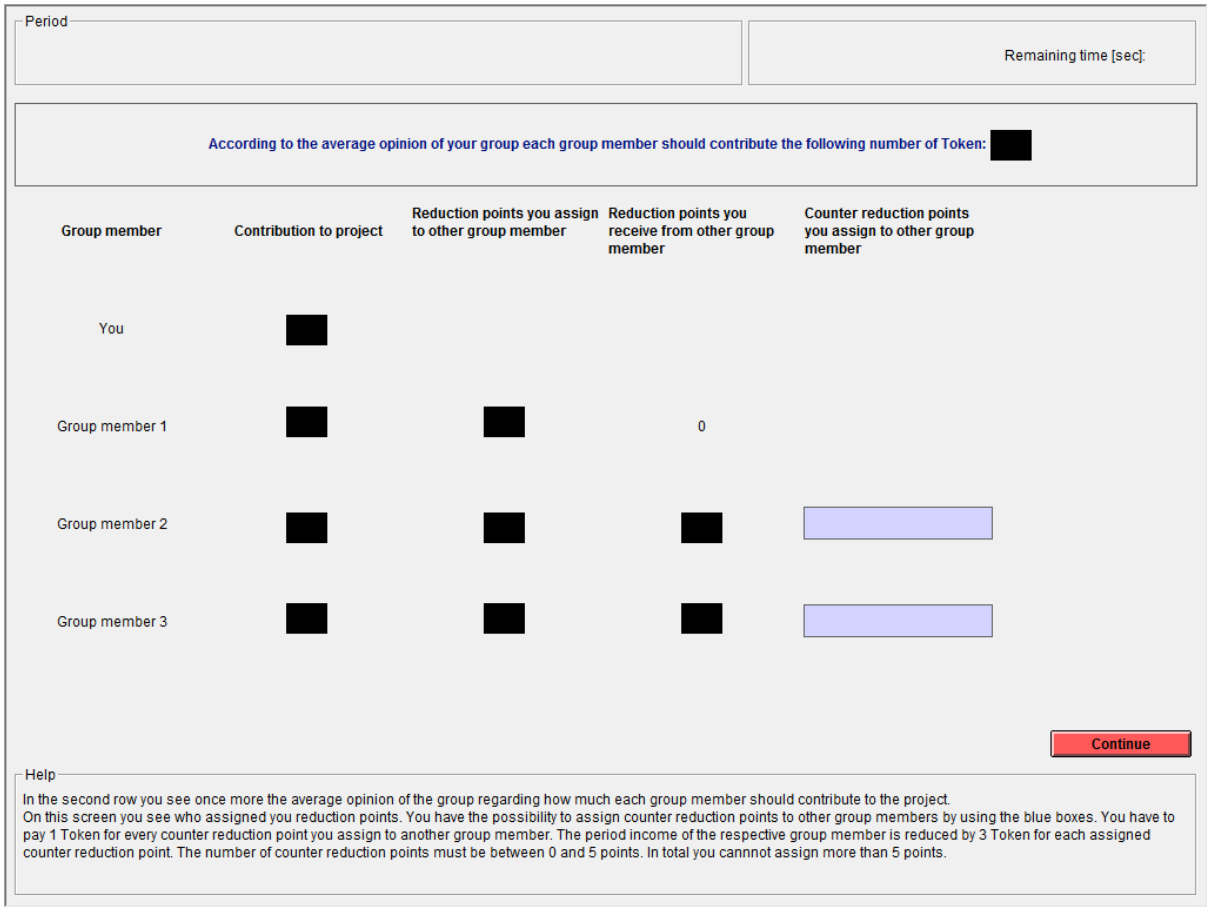
Screen stage 3: Input of the number of reduction points into the three empty boxes and confirmation with "Continue".

On the screen you will see, besides the indication of the period and the remaining time, how high the contribution of the other members was. **Your contribution** is indicated in **the first row** (labeled "you"). You will also see the contribution of the other members in the rows below. **Please be aware of the fact that the order in which the contributions of the other three group members are shown is different for each period, since the identification number for each group member will be randomly assigned every period**. More specifically, this means that the person behind the identification number 3 for example, can be a different group member from period to period. **Note the identification numbers stay the same within one period**.

Stage 4 – Assignment of counter reduction points

At the beginning of this stage each group member will receive an additional 5 Token. These Token can be used to reduce the earnings of those group members, who reduced

your own earnings in stage 3. During stage 4 you will receive a reminder on the monitor screen (see below) on how much each group member should contribute to the project according the average opinion. Furthermore, you will receive information about who assigned you how many reduction points and you will also see how many Tokens were contributed by you and by the other group members. Afterwards, you can state on the respective line if you want to assign counter reduction points. If that is the case, you have to state how many counter reduction points you want to assign. The cost of assigning a counter reduction point is, as in stage 3, 1 Token per point. Each received counter reduction point will lead to a reduction of earnings of 3 Token. You can distribute a maximum of 5 counter reduction points.



Screen stage 4: Input of your answer into the empty box and confirmation with "Continue"

Overview of total earnings in <u>one</u> period

At the end of each period you will be informed about your earnings during that period and how it is composed (see monitor screen below). You will see the contributions to the project of all group members, the number of reduction points that you assigned to the

other members, the number of received reduction points, the counter reduction points you assigned and your received counter reduction points. The formula below will show you how the earnings are composed during one period.

Earnings in one period of a group member

=

(20 Token) – (Contribution to the project)

+

(0.4 Token * sum of all contributions to the project)

+

(10 Token) – (1 Token * number of assigned reduction points)

-

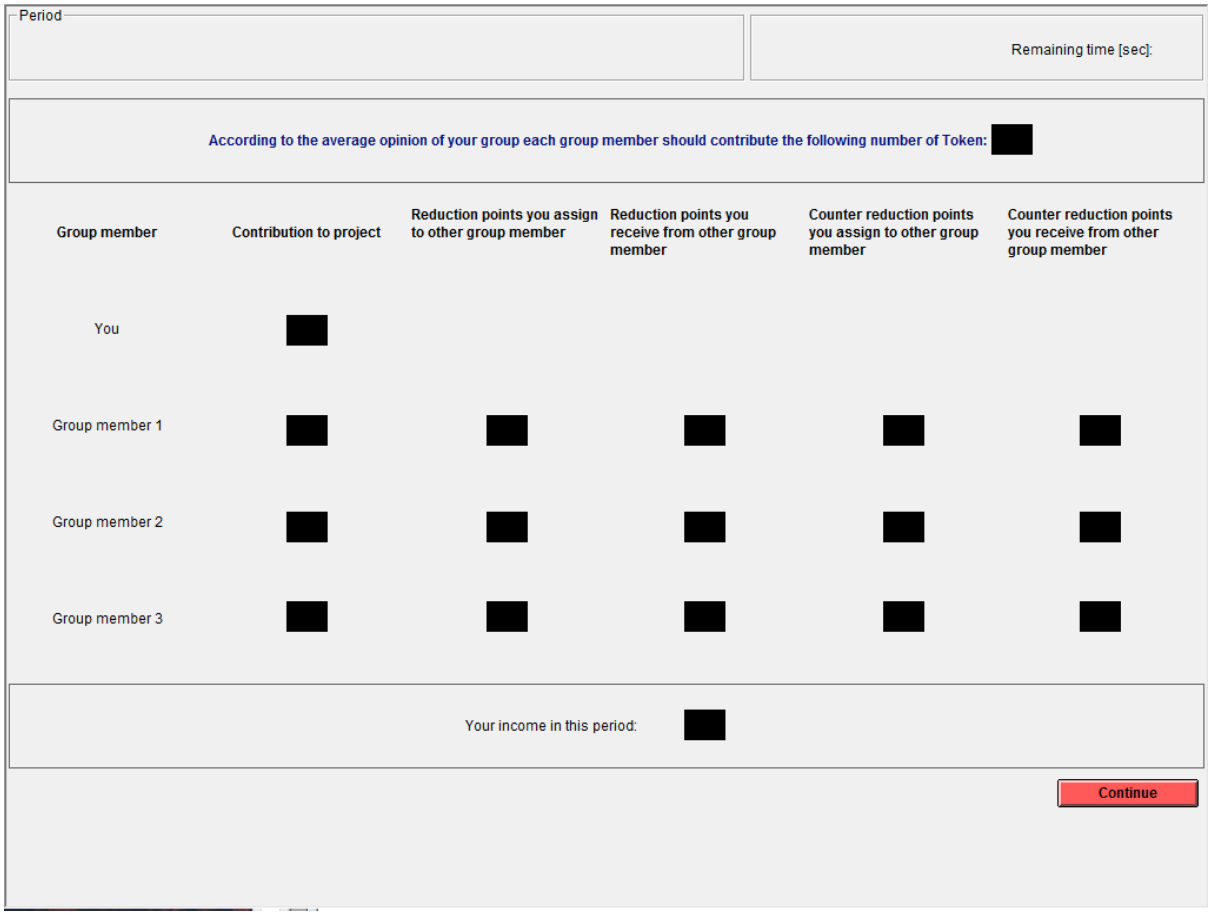
(3 Token * number of received reduction points)

+

(5 Token) - (1 Token * number of assigned counter reduction points)

-

(3 Tokens * number of received counter reduction points)



Screen at the end of each period showing an overview of the period and your period earnings. Confirmation with "Continue"

Control questions

Please answer the following questions and raise your hand as soon as you have finished.

1. Assuming nobody contributes anything (including you) to the project, and nobody assigns reduction points nor counter reduction points.

   How much are your earnings in this period? _____ Token

   How much are the earnings of the other group members? _____ Token

2. Assuming everybody contributes 20 Tokens to the project (including you). Furthermore, nobody assigns reduction points nor counter reduction points.

   How much are your earnings in this period? _____ Token

How much are the earnings of the other group members? _____ Token

3. Assuming the other three group members contribute in total 30 Token to the project. Furthermore, nobody assigns reduction points nor counter reduction points.

   a) How much are your earnings in this period if you contribute 0 Token (additionally to the 30 Token of the other members) to the project? _____ Token

   b) How much are your earnings in this period if you contribute 15 Tokens (additionally to the 30 Token of the other members) to the project?
   _____ Token

4. Assuming you contribute 8 Token to the project. Furthermore, nobody assigns reduction points nor counter reduction points.

   a) How much are your earnings if the other group members contribute in total 7 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

   b) How much are your earnings if the other group members contribute in total 22 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

5. Assuming you assign 5 reduction points to another group member.

   a) How much does this decrease the earnings of the other group member?
   _____ Token

   b) Assume another member assigns 2 counter reduction points to you. How much does this decrease your earnings?_____ Token

6. Can you assign counter reduction points during stage 4 to another member if you did not receive any reduction points from that member during stage 3?

   _____ YES _____ NO

7. Is the person with the identified as "group member 1" necessarily the same in each period?

   _____ YES _____ NO

Instructions sessions NF–noNF Part 2

Second part of the study

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

Procedure of the second part

The second part of the study consists again of 15 periods. Every period consists of the **same stages** as in the first part. The only exception is the **elimination of stage 1**. More precisely, this means that you and the other members **will not be asked** the following question at the beginning of the period: "In your opinion, how many Tokens should each group member contribute to the project?"

Stages 2, 3 and 4 stay the same as in the first part. Since stage 1 will not be a part of the experiment anymore, the following information will no longer be available: *"According to the average opinion of your group each group member should contribute the following number of Token:"*

You will form a group together with the same three participants as in the first part. The composition of the group will therefore **not** change.

Your total income will be the sum of your earnings in the first and second part.

Instructions sessions NFnoP–noNFnoP Part 1

Welcome

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

General information

You are now participating in an economic experiment. You will receive a fixed amount of 5 Pounds Sterling. During the study you will be able to earn more money. You will receive your earnings in cash at the end of the study.

During the experiment we talk about Token instead of Pounds Sterling. Initially, your earnings will therefore be calculated in Token. At the end of the experiment, the total sum of Token is converted into Pounds. The following condition will hold:

$$1 \text{ Token} = 1\text{p}$$

Every participant will get (additional to the show-up fee of £5) a one-time lump sum payment of **200 Token**. With this lump sum payment you will be able to cover possible losses. At the end of the experiment you will receive your total sum of Token (including the lump sum payment) in addition to the £4 show-up fee. Your earnings will be paid out in cash.

You are not allowed to communicate with the other participants. Please ask the experimenter if you have any questions. The violation of this rule will lead to the exclusion of the experiment and of all the above mentioned payments.

The data collected during the study will not be matched with your identity at any point.

Short description of the study

At the beginning of the experiment you will be randomly assigned to a group of four. Hence, there will be three other participants in the group with you. **The group composition will not change during the course of the study**. You will only interact with the members of your own group. Every group member has the same possibility as the other members and will receive the same instructions.

The experiment consists of two parts. The instructions for the second part will be handed out after the conclusion of the first part. Your total income will be a sum of the two parts. The first part of the experiment consists of 15 periods. Each of the 4 group members will have to decide in each period how many Token they want to contribute to the project. Each group member can contribute between 0 and 20 Token to the project. Each period consists of 2 stages:

1. During the first stage each group member has the opportunity to express how much he or she thinks that every group member should contribute to the project.

2. During the second stage each group member decides how many Token he or she will contribute to the project.

At the end of each period you will be informed about how much you will have earned during this period and about the composition of these earnings. On the following pages, we will describe the exact procedure of the experiment.

Procedure of the study

At the beginning of the first part you will be randomly assigned to a group of four. Hence, you and three other participants will together form one group. These groups will remain unchanged for the whole experiment.

At the beginning of each period – during stage 1 – you will be able to indicate how much each group member should contribute to the project in your opinion. The average value will be conveyed to all members of your group. During stage 2 you will decide how much you want to contribute to the project. You will receive a share of the earnings of the project, which in turn depends on the decisions of all group members. The earnings of the project will be divided equally among all four group members. Further details will be described below. During the first part of the experiment these 2 stages will be repeated 15 times. Afterwards the instructions for the second part will be distributed.

The stages of the experiment in detail

Stage 1 – Communication about how much each member should contribute to the project

At the beginning of each period you will be asked the following questions (see below the monitor screen for stage 1):

"In your opinion, how many Token should each group member contribute to the project?"

Since every group member can contribute between 0 and 20 Token to the project, you have to answer this question with a whole number from the range of 0 to 20.

Screen stage 1: Input of your answer into the empty box and confirmation with "Continue"

The average of the answers of your group will be calculated and subsequently conveyed to each group member. The average will be rounded to the nearest whole number. This information will be available on the monitor screen of stage 2 (see below).

Stage 2 – Decision about how much to contribute to the project

In every period each group member will get 20 Token. Each group member has to decide how much he or she wants to contribute to the project. Every whole number between 0 and 20 can be chosen. Each group member profits equally from the earnings of the project. **The sum of the contributed Token will be multiplied by 1.6 (+60% of all the contributions) and will be equally redistributed**.

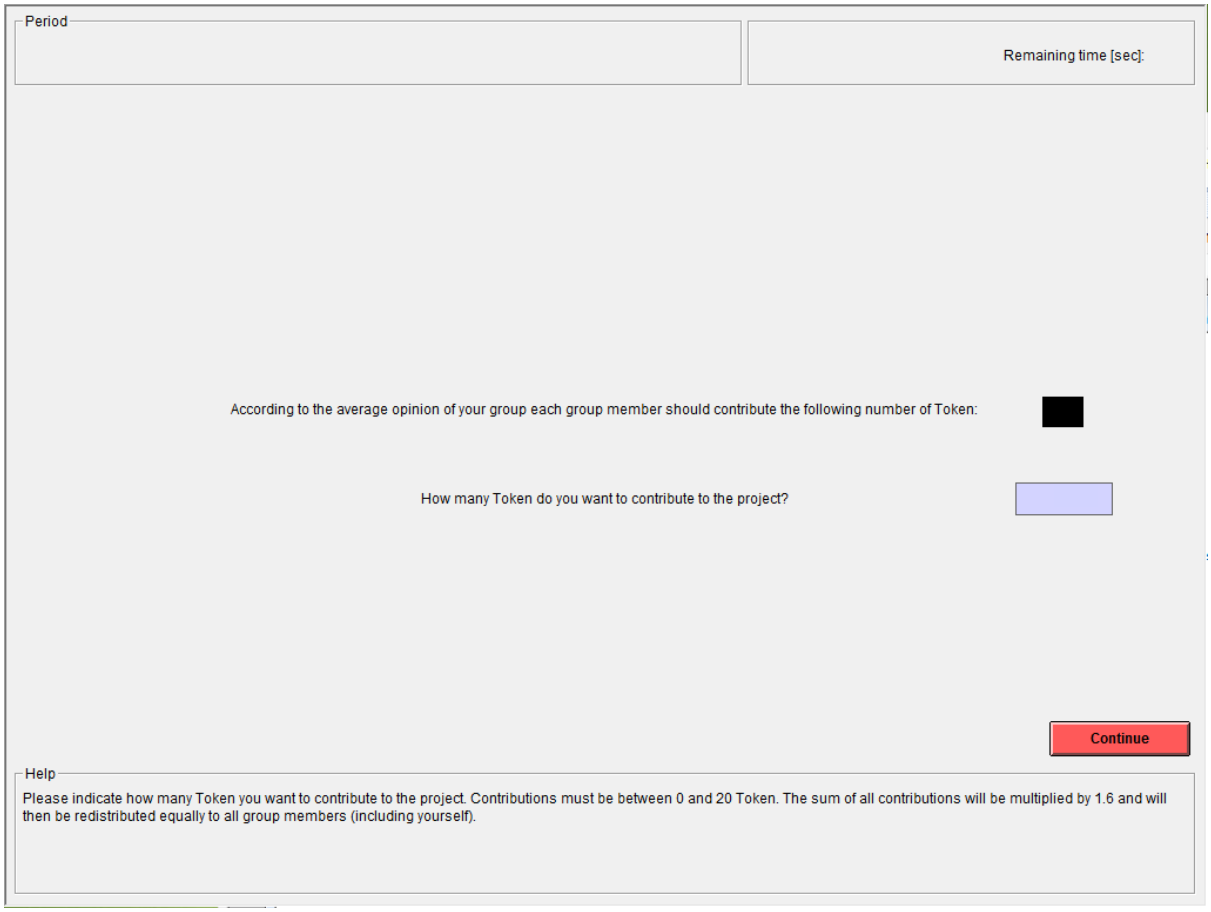
The earnings of the project can be calculated by 1.6*X Token, X Token being the sum of all contributions. This amount will be equally redistributed to all group members. In other words, each group member will receive $\frac{1.6*X}{4} = 0.4$ X Token. Therefore, for every

contributed Token, each group member will receive 0.4 Token (including you). You can keep the Token, which you will not have contributed, for yourself. These Token will be part of your total earnings. The input is made as shown in the monitor screen below.

Earnings from the project for each group member = 0.4 * amount of contributed Tokens

Examples:

- Assuming each group member will contribute 20 Token to the project, 80 Token will be available for the project in total. Each group member will receive 0.4*80 = 32 Token from the project.
- Assuming nobody will contribute to the project (0 Token), then nobody will receive earnings from the project since every group member decided to keep 20 Token for him- or herself.
- Assuming you contribute 5 Token to the project and each of the other group members contributes 10 Token, then you will get 0.4*(5+10+10+10) = 14 Token in addition to the 15 Token (which you did not contribute). The other group members will get 14 Token from the project as well, but they will only have 10 Token left from before.



Screen stage 2: Input of your answer into the empty box and confirmation with "Continue"

Overview of total earnings in one period

At the end of each period you will be informed about your earnings during that period and how it is composed (see monitor screen below). **Your contribution** is indicated in the **first row** (labeled "you"). You will also see the contribution of the other members in the rows below. **Please be aware of the fact that the order in which the contributions of the other three group members are shown is different for each period, since the identification number for each group member will be randomly assigned every period**. More specifically, this means that the person behind the identification number 3 for example, can be a different group member from period to period. **Note the identification numbers stay the same within one period**. At the end of each period each group member receives an additional 15 Token, these 15 Token are part of your period income. The formula below will show you how the earnings are composed during one period.

Earnings in one period of a group member

=

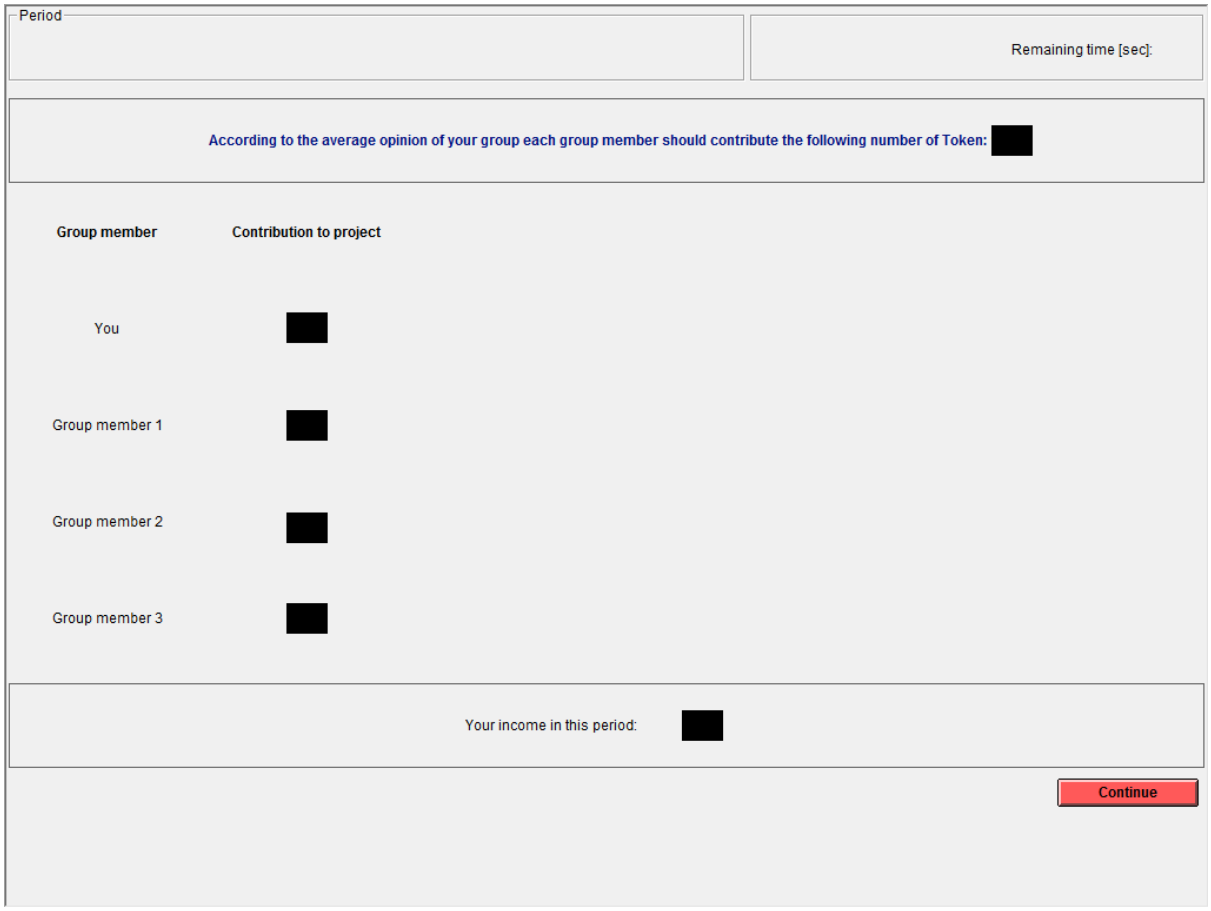
(20 Token) – (Contribution to the project)

+

(0.4 Token * sum of all contributions to the project)

+

(15 Token)



Screen at the end of each period showing an overview of the period and your period earnings. Confirmation with "Continue"

Control questions

Please answer the following questions and raise your hand as soon as you have finished.

1. Assuming nobody contributes anything (including you) to the project.

   How much are your earnings in this period? _____ Token

   How much are the earnings of the other group members? _____ Token

2. Assuming everybody contributes 20 Tokens to the project (including you).

   How much are your earnings in this period? _____ Token

   How much are the earnings of the other group members? _____ Token

3. Assuming the other three group members contribute in total 30 Token to the project.

   a) How much are your earnings in this period if you contribute 0 Token (additionally to the 30 Token of the other members) to the project? _____ Token

   b) How much are your earnings in this period if you contribute 15 Tokens (additionally to the 30 Token of the other members) to the project?
   _____ Token

4. Assuming you contribute 8 Token to the project.

   a) How much are your earnings if the other group members contribute in total 7 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

   b) How much are your earnings if the other group members contribute in total 22 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

5. Is the person with the identified as "group member 1" necessarily the same in each period?

   ____ YES ____ NO

Instructions sessions NFnoP–noNFnoP Part 2

Second part of the study

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

Procedure of the second part

The second part of the study consists again of 15 periods. Stage 2 remains the same as in the first part. However, **stage 1 is eliminated**. More precisely, this means that you and the other members **will not be asked** the following question at the beginning of the period: "In your opinion, how many Tokens should each group member contribute to the project?"

Stage 2 stay the same as in the first part. Since stage 1 will not be a part of the experiment anymore, the following information will no longer be available: *"According to the average opinion of your group each group member should contribute the following number of Token:"*

You will form a group together with the same three participants as in the first part. The composition of the group will therefore **not** change.

Your total income will be the sum of your earnings in the first and second part.

Instructions sessions noNF–NF Part 1

Welcome

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

General information

You are now participating in an economic experiment. You will receive a fixed amount of 5 Pounds Sterling. During the study you will be able to earn more money. You will receive your earnings in cash at the end of the study.

During the experiment we talk about Token instead of Pounds Sterling. Initially, your earnings will therefore be calculated in Token. At the end of the experiment, the total sum of Token is converted into Pounds. The following condition will hold:

$$1 \text{ Token} = 1\text{p}$$

Every participant will get (additional to the show-up fee of £5) a one-time lump sum payment of **200 Token**. With this lump sum payment you will be able to cover possible losses. At the end of the experiment you will receive your total sum of Token (including the lump sum payment) in addition to the £4 show-up fee. Your earnings will be paid out in cash.

You are not allowed to communicate with the other participants. Please ask the experimenter if you have any questions. The violation of this rule will lead to the exclusion of the experiment and of all the above mentioned payments.

The data collected during the study will not be matched with your identity at any point.

Short description of the study

At the beginning of the experiment you will be randomly assigned to a group of four. Hence, there will be three other participants in the group with you. **The group composition will not change during the course of the study**. You will only interact with the members of your own group. Every group member has the same possibility as the other members and will receive the same instructions.

The experiment consists of two parts. The instructions for the second part will be handed out after the conclusion of the first part. Your total income will be a sum of the two parts. The first part of the experiment consists of 15 periods. Each of the 4 group members will have to decide in each period how many Token they want to contribute to the project. Each group member can contribute between 0 and 20 Token to the project. Each period consists of 3 stages:

1. During the first stage each group member decides how many Token he or she will contribute to the project.

2. During the second stage every group member will be informed about how many Token will have been contributed by the other group members. Afterwards the

members will be able to spend Token in order to reduce the earnings of the other group members.

3. During the final stage the group members will again get the chance to spend Token in order to reduce the earnings of the other group members. They will, however, only be able to reduce the earnings of those group members, who reduced their earnings during the second stage.

At the end of each period you will be informed about how much you will have earned during this period and about the composition of these earnings. On the following pages, we will describe the exact procedure of the experiment.

Procedure of the study

At the beginning of the first part you will be randomly assigned to a group of four. Hence, you and three other participants will together form one group. These groups will remain unchanged for the whole experiment.

At the beginning of each period – during stage 1 – you will decide how much you want to contribute to the project. You will receive a share of the earnings of the project, which in turn depends on the decisions of all group members. The earnings of the project will be divided equally among all four group members. Further details will be described below. During stage 2 you will be informed about how much the other group members will indeed have contributed to the project. In addition, you will be able to use your Token in order to reduce the earnings of the other group members during this stage. This will be possible through the assignment of reduction points. During stage 3 the members whose earnings were reduced by other group members during stage 2, will in turn be able to assign counter reduction points to those and only those group members. During the first part of the experiment these 3 stages will be repeated 15 times. Afterwards the instructions for the second part will be distributed.

The stages of the experiment in detail

Stage 1 – Decision about how much to contribute to the project

In every period each group member will get 20 Token. Each group member has to decide how much he or she wants to contribute to the project. Every whole number between

0 and 20 can be chosen. Each group member profits equally from the earnings of the project. **The sum of the contributed Token will be multiplied by 1.6 (+60% of all the contributions) and will be equally redistributed**.

The earnings of the project can be calculated by 1.6*X Token, X Token being the sum of all contributions. This amount will be equally redistributed to all group members. In other words, each group member will receive $\frac{1.6*X}{4}$ = 0.4 X Token. Therefore, for every contributed Token, each group member will receive 0.4 Token (including you). You can keep the Token, which you will not have contributed, for yourself. These Token will be part of your total earnings. The input is made as shown in the monitor screen below.

Earnings from the project for each group member
=
0.4 * amount of contributed Tokens

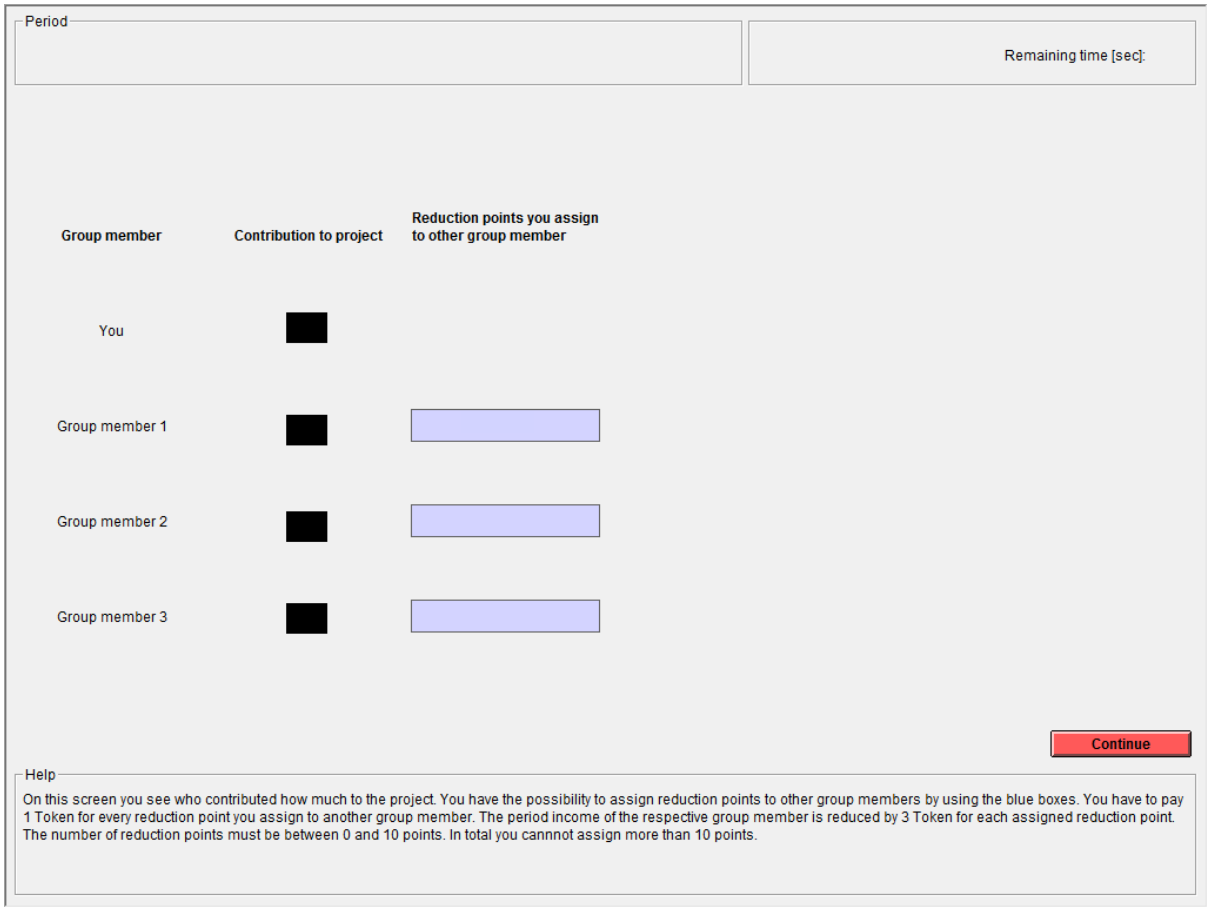
Examples:

- Assuming each group member will contribute 20 Token to the project, 80 Token will be available for the project in total. Each group member will receive 0.4*80 = 32 Token from the project.
- Assuming nobody will contribute to the project (0 Token), then nobody will receive earnings from the project since every group member decided to keep 20 Token for him- or herself.
- Assuming you contribute 5 Token to the project and each of the other group members contributes 10 Token, then you will get 0.4*(5+10+10+10) = 14 Token in addition to the 15 Token (which you did not contribute). The other group members will get 14 Token from the project as well, but they will only have 10 Token left from before.

Screen stage 1: Input of your answer into the empty box and confirmation with "Continue"

Stage 2 – Assignment of reduction points

At the beginning of this stage each group member will receive 10 additional Token. These Token can be used to reduce the earnings of the other members of the group. You can do this by assigning reduction points. At the beginning of this stage you will see how much the other group members will have contributed to the project (see monitor screen below). Afterwards, you will decide whether or not you want to assign reduction points to other group members, and in case you want to do so, how many reduction points you want to assign. You have to pay 1 Token for each reduction point you want to assign. The group member's earnings will be reduced by 3 Token for every reduction point received. More specifically, you can pay 1 Token to reduce the earnings of another member by 3 Tokens. You can distribute a maximum of 10 reduction points. The other members have the same possibility as you do.



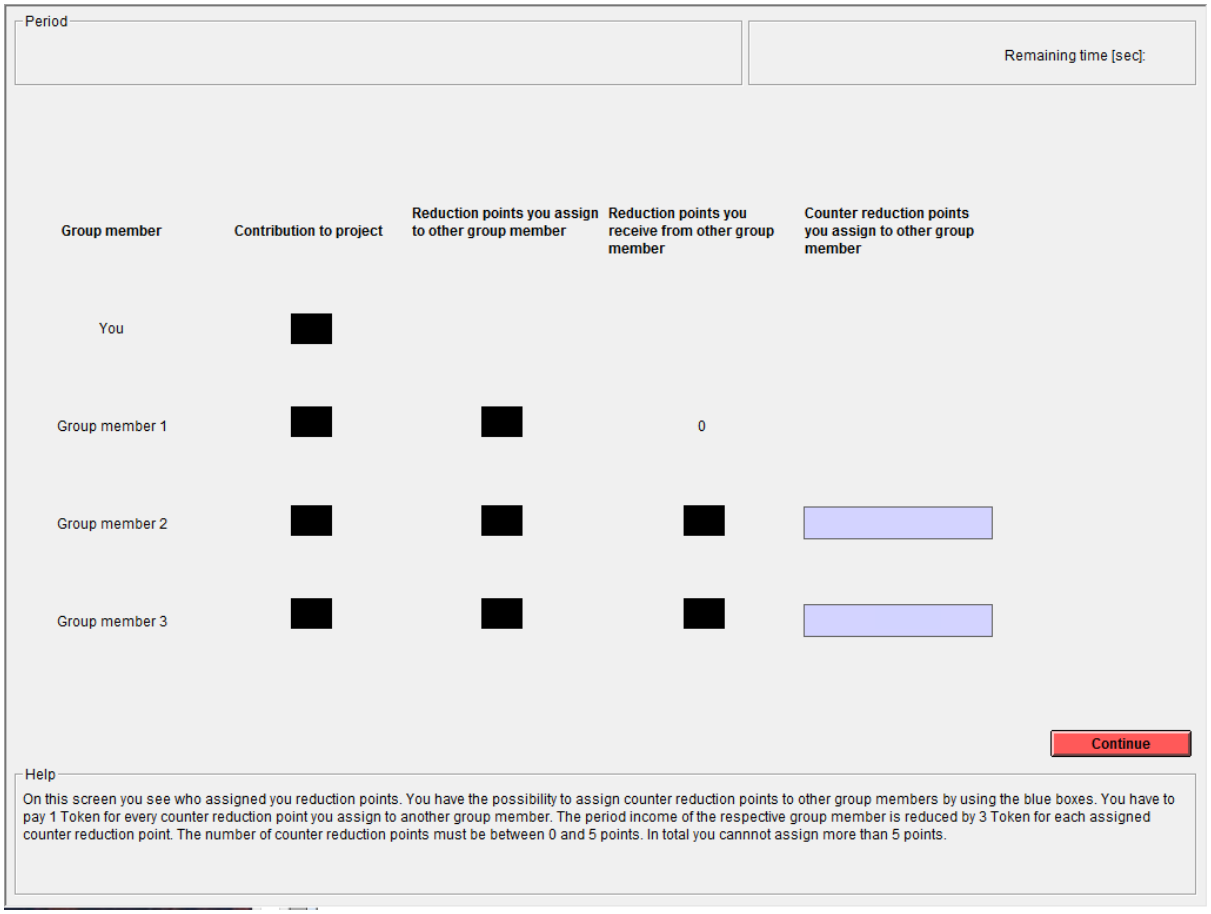
Screen stage 2: Input of the number of reduction points into the three empty boxes and confirmation with "Continue".

On the screen you will see, besides the indication of the period and the remaining time, how high the contribution of the other members was. **Your contribution** is indicated in **the first row** (labeled "you"). You will also see the contribution of the other members in the rows below. **Please be aware of the fact that the order in which the contributions of the other three group members are shown is different for each period, since the identification number for each group member will be randomly assigned every period**. More specifically, this means that the person behind the identification number 3 for example, can be a different group member from period to period. **Note the identification numbers stay the same within one period**.

Stage 3 – Assignment of counter reduction points

At the beginning of this stage each group member will receive an additional 5 Token. These Token can be used to reduce the earnings of those group members, who reduced

your own earnings in stage 2. Furthermore, you will receive information about who assigned you how many reduction points and you will also see how many Tokens were contributed by you and by the other group members. Afterwards, you can state on the respective line if you want to assign counter reduction points. If that is the case, you have to state how many counter reduction points you want to assign. The cost of assigning a counter reduction point is, as in stage 2, 1 Token per point. Each received counter reduction point will lead to a reduction of earnings of 3 Token. You can distribute a maximum of 5 counter reduction points.



Screen stage 3: Input of your answer into the empty box and confirmation with "Continue"

Overview of total earnings in one period

At the end of each period you will be informed about your earnings during that period and how it is composed (see monitor screen below). You will see the contributions to the project of all group members, the number of reduction points that you assigned to the other members, the number of received reduction points, the counter reduction points you assigned and your received counter reduction points. The formula below will show

you how the earnings are composed during one period.

Earnings in one period of a group member

=

(20 Token) – (Contribution to the project)

+

(0.4 Token * sum of all contributions to the project)

+

(10 Token) – (1 Token * number of assigned reduction points)

-

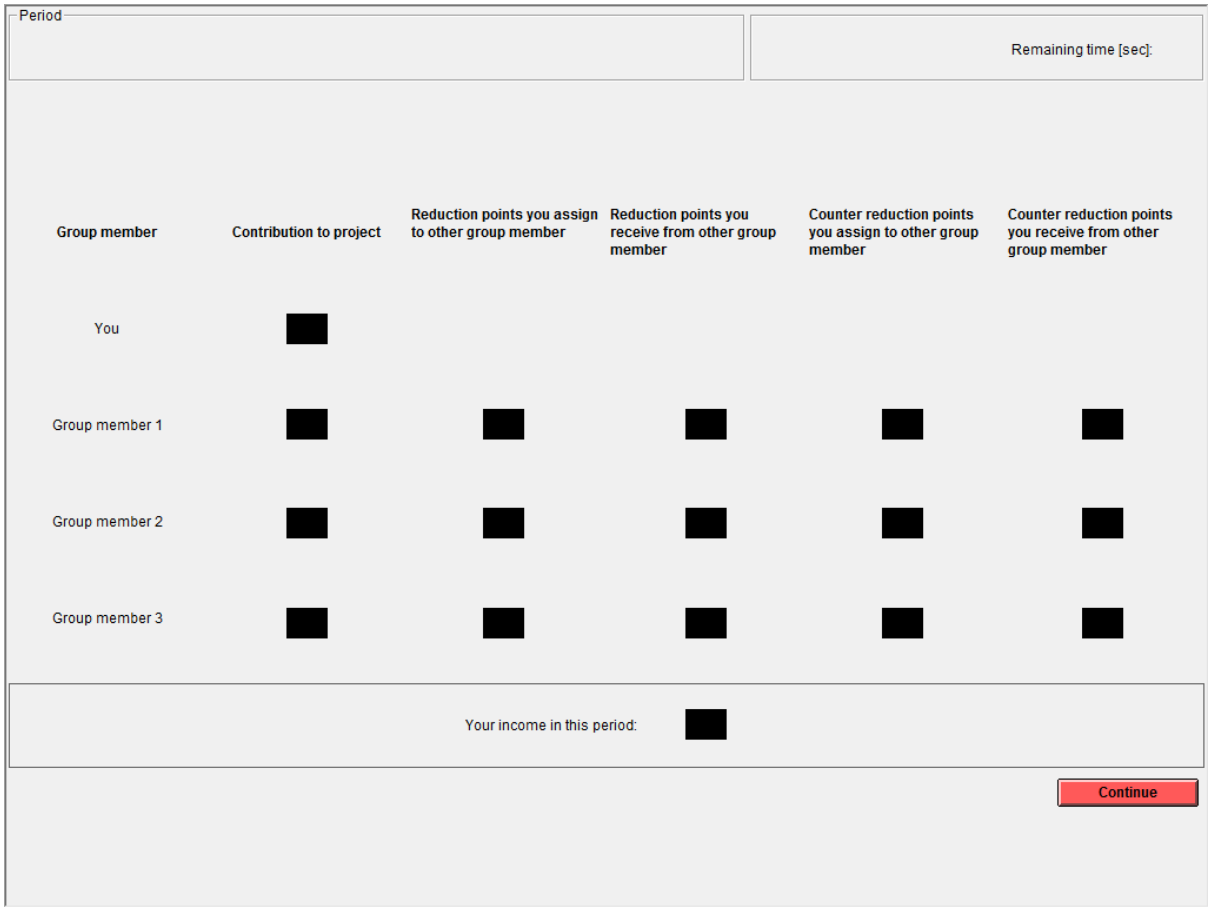
(3 Token * number of received reduction points)

+

(5 Token) - (1 Token * number of assigned counter reduction points)

-

(3 Tokens * number of received counter reduction points)



Screen at the end of each period showing an overview of the period and your period earnings. Confirmation with "Continue"

Control questions

Please answer the following questions and raise your hand as soon as you have finished.

1. Assuming nobody contributes anything (including you) to the project, and nobody assigns reduction points nor counter reduction points.

   How much are your earnings in this period? _____ Token

   How much are the earnings of the other group members? _____ Token

2. Assuming everybody contributes 20 Tokens to the project (including you). Furthermore, nobody assigns reduction points nor counter reduction points.

   How much are your earnings in this period? _____ Token

How much are the earnings of the other group members? _____ Token

3. Assuming the other three group members contribute in total 30 Token to the project. Furthermore, nobody assigns reduction points nor counter reduction points.

   a) How much are your earnings in this period if you contribute 0 Token (additionally to the 30 Token of the other members) to the project? _____ Token

   b) How much are your earnings in this period if you contribute 15 Tokens (additionally to the 30 Token of the other members) to the project?
   _____ Token

4. Assuming you contribute 8 Token to the project. Furthermore, nobody assigns reduction points nor counter reduction points.

   a) How much are your earnings if the other group members contribute in total 7 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

   b) How much are your earnings if the other group members contribute in total 22 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

5. Assuming you assign 5 reduction points to another group member.

   a) How much does this decrease the earnings of the other group member?
   _____ Token

   b) Assume another member assigns 2 counter reduction points to you. How much does this decrease your earnings?_____ Token

6. Can you assign counter reduction points during stage 3 to another member if you did not receive any reduction points from that member during stage 2?

   _____ YES _____ NO

7. Is the person with the identified as "group member 1" necessarily the same in each period?

   _____ YES _____ NO

Instructions sessions noNF–NF Part 2

Second part of the study

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

Procedure of the second part

The second part of the study consists again of 15 periods. Every period consists of the **same 3 stages** as in the first part. The only change will be the incorporation of a new stage at the beginning of each period. This new stage will be implemented before stage 1 from the first part. During this stage each member will have the possibility to say how much he or she thinks that each group member should contribute to the project.

New stage – Communication about how much each member should contribute to the project

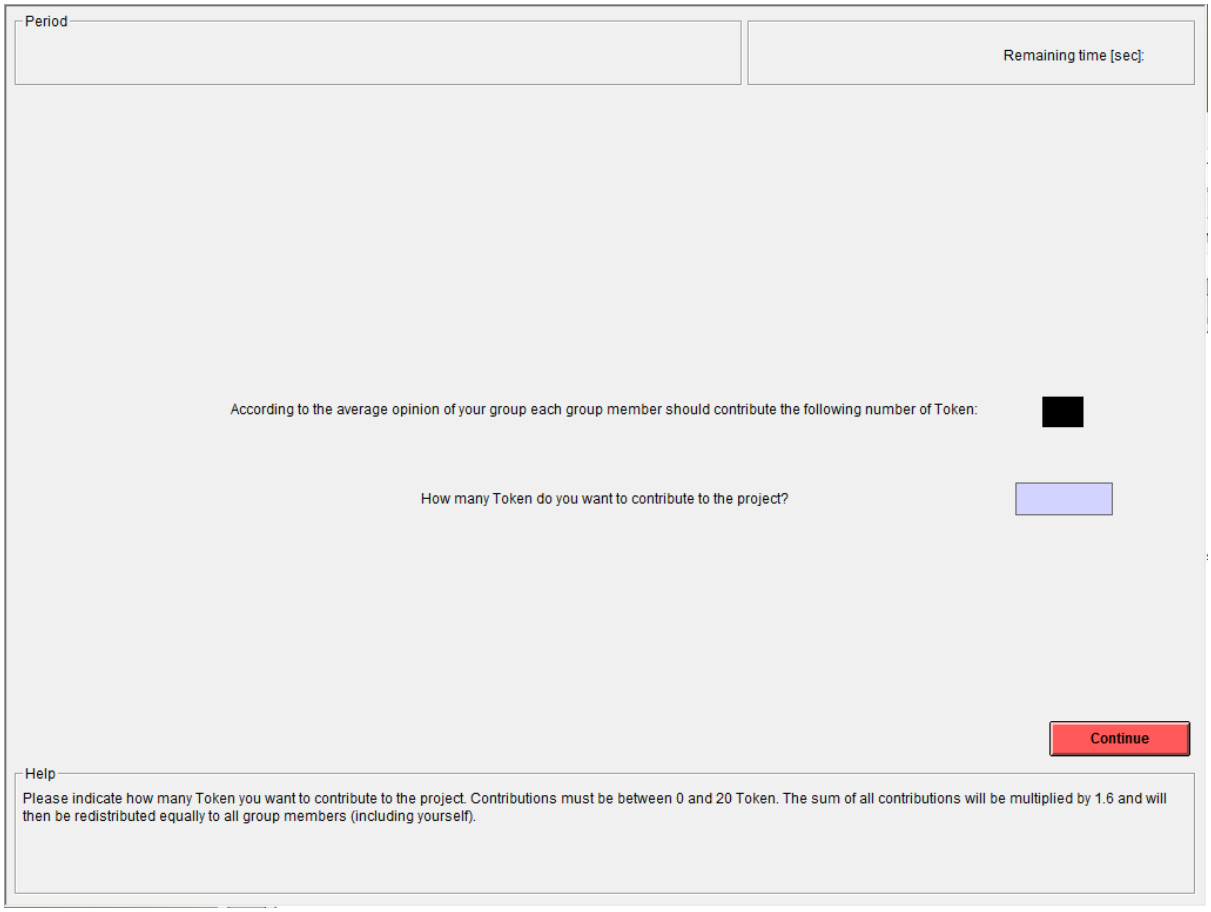
At the beginning of each period you will be asked the following questions (see below the monitor screen for new stage):

"In your opinion, how many Token should each group member contribute to the project?"

Since every group member can contribute between 0 and 20 Token to the project, you have to answer this question with a whole number from the range of 0 to 20.

Screen new stage: Input of your answer into the empty box and confirmation with "Continue"

The average of the answers of your group will be calculated and subsequently conveyed to each group member. The average will be rounded to the nearest whole number. This information will be available on the monitor screen of the stage during which you will have to decide how much you want to contribute to the project (see below).



New monitor screen for the stage – decision of contribution to the project: Input of your answer into the empty box and confirmation with "Continue"

Stage 1, 2 and 3 will stay the same as in the first part. Since we are incorporating a new stage the subsequent information will be available during all of the following stages: "According to the average opinion of the group each group member should contribute the following number of Token:". This information will be available in the header during stage 2 (assignment of reduction points), stage 3 (assignment of counter reduction points), as well as during the overview of the period earnings.

You will form a group together with the same three participants as in the first part. The composition of the group will therefore **not** change.

Your total income will be the sum of your earnings in the first and second part.

Instructions sessions noNFnoP–NFnoP Part 1

Welcome

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

General information

You are now participating in an economic experiment. You will receive a fixed amount of 5 Pounds Sterling. During the study you will be able to earn more money. You will receive your earnings in cash at the end of the study.

During the experiment we talk about Token instead of Pounds Sterling. Initially, your earnings will therefore be calculated in Token. At the end of the experiment, the total sum of Token is converted into Pounds. The following condition will hold:

$$1 \text{ Token} = 1\text{p}$$

Every participant will get (additional to the show-up fee of £5) a one-time lump sum payment of **200 Token**. With this lump sum payment you will be able to cover possible losses. At the end of the experiment you will receive your total sum of Token (including the lump sum payment) in addition to the £4 show-up fee. Your earnings will be paid out in cash.

You are not allowed to communicate with the other participants. Please ask the experimenter if you have any questions. The violation of this rule will lead to the exclusion of the experiment and of all the above mentioned payments.

The data collected during the study will not be matched with your identity at any point.

Short description of the study

At the beginning of the experiment you will be randomly assigned to a group of four. Hence, there will be three other participants in the group with you. **The group composition will not change during the course of the study**. You will only interact with the members of your own group. Every group member has the same possibility as

the other members and will receive the same instructions.

The experiment consists of two parts. The instructions for the second part will be handed out after the conclusion of the first part. Your total income will be a sum of the two parts. The first part of the experiment consists of 15 periods. Each of the 4 group members will have to decide in each period how many Token they want to contribute to the project. Each group member can contribute between 0 and 20 Token to the project. Each period consists of 1 stage:

1. During this stage each group member decides how many Token he or she will contribute to the project.

At the end of each period you will be informed about how much you will have earned during this period and about the composition of these earnings. On the following pages, we will describe the exact procedure of the experiment.

Procedure of the study

At the beginning of the first part you will be randomly assigned to a group of four. Hence, you and three other participants will together form one group. These groups will remain unchanged for the whole experiment.

In each stage you will decide how much you want to contribute to the project. You will receive a share of the earnings of the project, which in turn depends on the decisions of all group members. The earnings of the project will be divided equally among all four group members. Further details will be described below. During the first part of the experiment such a period will be repeated 15 times. Afterwards the instructions for the second part will be distributed.

The experiment in detail

Decision about how much to contribute to the project

In every period each group member will get 20 Token. Each group member has to decide how much he or she wants to contribute to the project. Every whole number between 0 and 20 can be chosen. Each group member profits equally from the earnings of the

project. **The sum of the contributed Token will be multiplied by 1.6 (+60% of all the contributions) and will be equally redistributed**.

The earnings of the project can be calculated by 1.6*X Token, X Token being the sum of all contributions. This amount will be equally redistributed to all group members. In other words, each group member will receive $\frac{1.6*X}{4}$ = 0.4 X Token. Therefore, for every contributed Token, each group member will receive 0.4 Token (including you). You can keep the Token, which you will not have contributed, for yourself. These Token will be part of your total earnings. The input is made as shown in the monitor screen below.

Earnings from the project for each group member = 0.4 * amount of contributed Tokens

Examples:

- Assuming each group member will contribute 20 Token to the project, 80 Token will be available for the project in total. Each group member will receive 0.4*80 = 32 Token from the project.
- Assuming nobody will contribute to the project (0 Token), then nobody will receive earnings from the project since every group member decided to keep 20 Token for him- or herself.
- Assuming you contribute 5 Token to the project and each of the other group members contributes 10 Token, then you will get 0.4*(5+10+10+10) = 14 Token in addition to the 15 Token (which you did not contribute). The other group members will get 14 Token from the project as well, but they will only have 10 Token left from before.

Screen contribution: Input of your answer into the empty box and confirmation with "Continue"

Overview of total earnings in one period

At the end of each period you will be informed about your earnings during that period and how it is composed (see monitor screen below). **Your contribution** is indicated in the **first row** (labeled "you"). You will also see the contribution of the other members in the rows below. **Please be aware of the fact that the order in which the contributions of the other three group members are shown is different for each period, since the identification number for each group member will be randomly assigned every period**. More specifically, this means that the person behind the identification number 3 for example, can be a different group member from period to period. **Note the identification numbers stay the same within one period**. At the end of each period each group member receives an additional 15 Token, these 15 Token are part of your period income. The formula below will show you how the earnings are composed during one period.

Earnings in one period of a group member

=

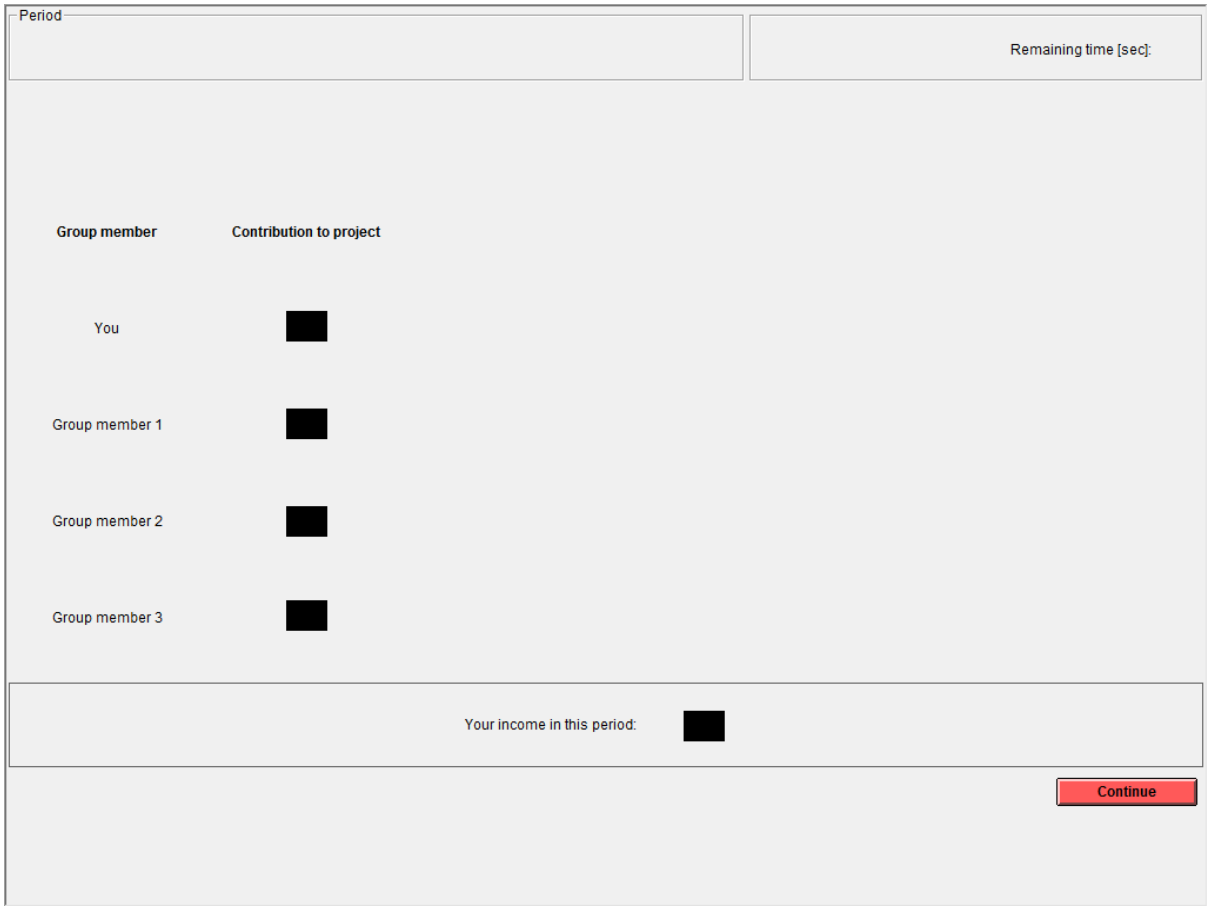
(20 Token) – (Contribution to the project)

+

(0.4 Token * sum of all contributions to the project)

+

(15 Token)



Screen at the end of each period showing an overview of the period and your period earnings. Confirmation with "Continue"

Control questions

Please answer the following questions and raise your hand as soon as you have finished.

1. Assuming nobody contributes anything (including you) to the project.

   How much are your earnings in this period? _____ Token

   How much are the earnings of the other group members? _____ Token

2. Assuming everybody contributes 20 Tokens to the project (including you).

   How much are your earnings in this period? _____ Token

   How much are the earnings of the other group members? _____ Token

3. Assuming the other three group members contribute in total 30 Token to the project.

   a) How much are your earnings in this period if you contribute 0 Token (additionally to the 30 Token of the other members) to the project? _____ Token

   b) How much are your earnings in this period if you contribute 15 Tokens (additionally to the 30 Token of the other members) to the project?
   _____ Token

4. Assuming you contribute 8 Token to the project.

   a) How much are your earnings if the other group members contribute in total 7 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

   b) How much are your earnings if the other group members contribute in total 22 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

5. Is the person with the identified as "group member 1" necessarily the same in each period?

   ____ YES ____ NO

Instructions sessions noNFnoP–NFnoP Part 2

The second part of the study consists again of 15 periods. Every period consists of the **same contribution decision** as in the first part. **Additionally, there is a new stage at right at the beginning for every period** (before the decision about the contribution). During this stage each member will have the possibility to say how much he or she thinks that each group member should contribute to the project.

New stage – Communication about how much each member should contribute to the project

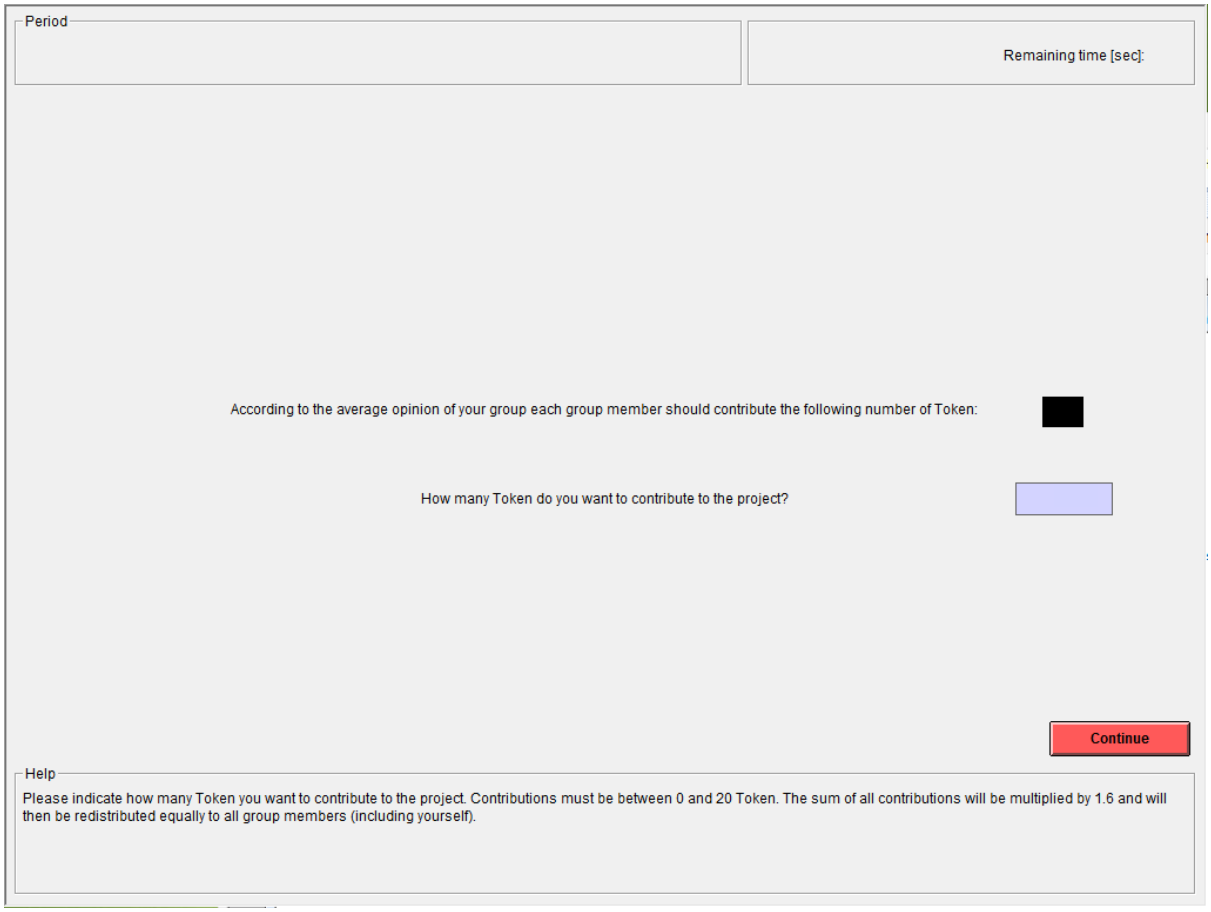
At the beginning of each period you will be asked the following questions (see below the monitor screen for new stage):

"In your opinion, how many Token should each group member contribute to the project?"

Since every group member can contribute between 0 and 20 Token to the project, you have to answer this question with a whole number from the range of 0 to 20.

Screen new stage: Input of your answer into the empty box and confirmation with "Continue"

The average of the answers of your group will be calculated and subsequently conveyed to each group member. The average will be rounded to the nearest whole number. This information will be available on the monitor screen of the stage during which you will have to decide how much you want to contribute to the project (see below).



New monitor screen for the stage – decision of contribution to the project: Input of your answer into the empty box and confirmation with "Continue"

You will form a group together with the same three participants as in the first part. The composition of the group will therefore **not** change.

Your total income will be the sum of your earnings in the first and second part.

Instructions sessions noNF–noNF Part 1

Welcome

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

General information

You are now participating in an economic experiment. You will receive a fixed amount of 5 Pounds Sterling. During the study you will be able to earn more money. You will receive your earnings in cash at the end of the study.

During the experiment we talk about Token instead of Pounds Sterling. Initially, your earnings will therefore be calculated in Token. At the end of the experiment, the total sum of Token is converted into Pounds. The following condition will hold:

$$1 \text{ Token} = 1\text{p}$$

Every participant will get (additional to the show-up fee of £5) a one-time lump sum payment of **200 Token**. With this lump sum payment you will be able to cover possible losses. At the end of the experiment you will receive your total sum of Token (including the lump sum payment) in addition to the £4 show-up fee. Your earnings will be paid out in cash.

You are not allowed to communicate with the other participants. Please ask the experimenter if you have any questions. The violation of this rule will lead to the exclusion of the experiment and of all the above mentioned payments.

The data collected during the study will not be matched with your identity at any point.

Short description of the study

At the beginning of the experiment you will be randomly assigned to a group of four. Hence, there will be three other participants in the group with you. **The group composition will not change during the course of the study**. You will only interact with the members of your own group. Every group member has the same possibility as the other members and will receive the same instructions.

The experiment consists of two parts. The instructions for the second part will be handed out after the conclusion of the first part. Your total income will be a sum of the two parts. The first part of the experiment consists of 15 periods. Each of the 4 group members will have to decide in each period how many Token they want to contribute to the project. Each group member can contribute between 0 and 20 Token to the project. Each period consists of 3 stages:

1. During the first stage each group member decides how many Token he or she will contribute to the project.

2. During the second stage every group member will be informed about how many Token will have been contributed by the other group members. Afterwards the members will be able to spend Token in order to reduce the earnings of the other group members.

3. During the final stage the group members will again get the chance to spend Token in order to reduce the earnings of the other group members. They will, however, only be able to reduce the earnings of those group members, who reduced their earnings during the second stage.

At the end of each period you will be informed about how much you will have earned during this period and about the composition of these earnings. On the following pages, we will describe the exact procedure of the experiment.

Procedure of the study

At the beginning of the first part you will be randomly assigned to a group of four. Hence, you and three other participants will together form one group. These groups will remain unchanged for the whole experiment.

At the beginning of each period – during stage 1 – you will decide how much you want to contribute to the project. You will receive a share of the earnings of the project, which in turn depends on the decisions of all group members. The earnings of the project will be divided equally among all four group members. Further details will be described below. During stage 2 you will be informed about how much the other group members will indeed have contributed to the project. In addition, you will be able to use your Token in order to reduce the earnings of the other group members during this stage. This will be possible through the assignment of reduction points. During stage 3 the members whose earnings were reduced by other group members during stage 2, will in turn be able to assign counter reduction points to those and only those group members. During the first part of the experiment these 3 stages will be repeated 15 times. Afterwards the instructions for the second part will be distributed.

The stages of the experiment in detail

Stage 1 – Decision about how much to contribute to the project

In every period each group member will get 20 Token. Each group member has to decide how much he or she wants to contribute to the project. Every whole number between 0 and 20 can be chosen. Each group member profits equally from the earnings of the project. **The sum of the contributed Token will be multiplied by 1.6 (+60% of all the contributions) and will be equally redistributed**.

The earnings of the project can be calculated by 1.6*X Token, X Token being the sum of all contributions. This amount will be equally redistributed to all group members. In other words, each group member will receive $\frac{1.6*X}{4}$ = 0.4 X Token. Therefore, for every contributed Token, each group member will receive 0.4 Token (including you). You can keep the Token, which you will not have contributed, for yourself. These Token will be part of your total earnings. The input is made as shown in the monitor screen below.

Earnings from the project for each group member
=
0.4 * amount of contributed Tokens

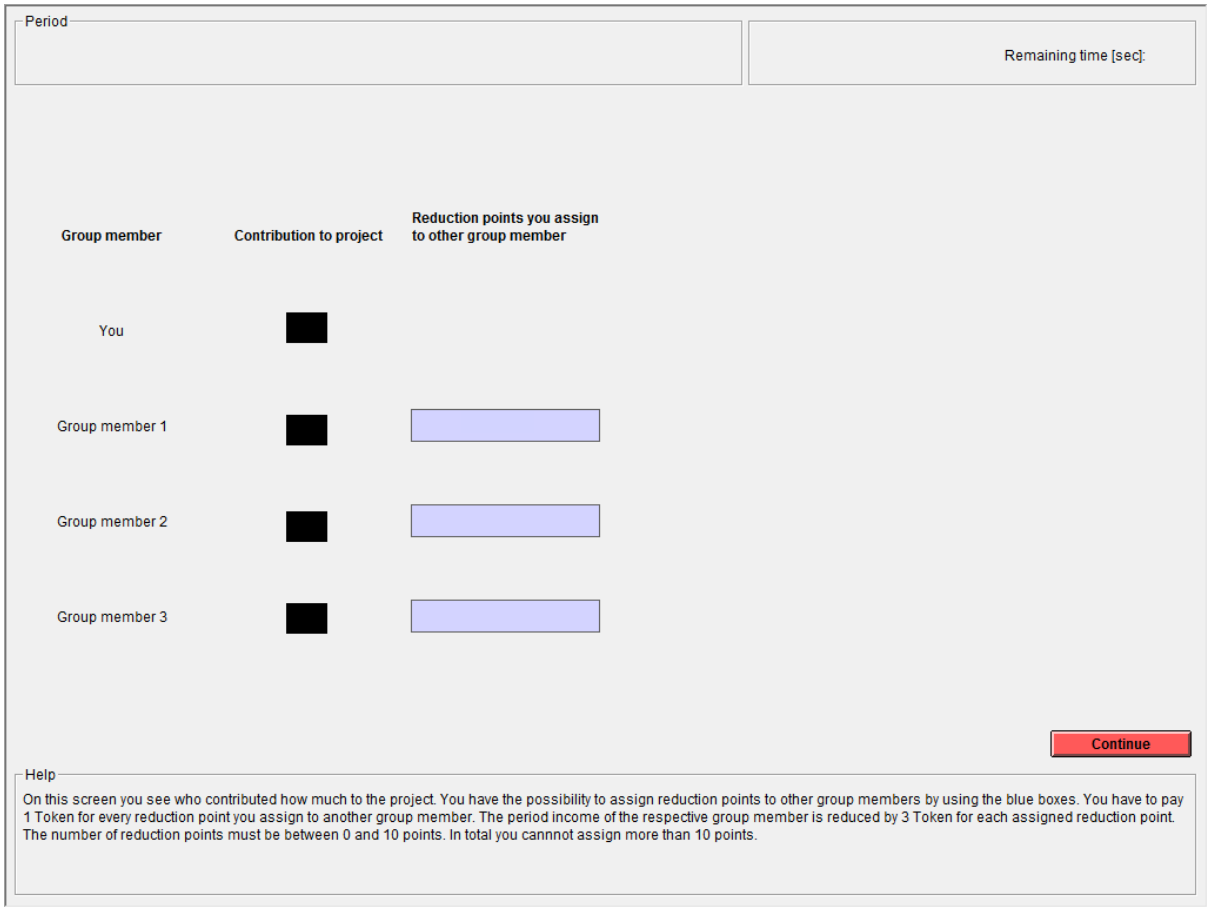
Examples:

- Assuming each group member will contribute 20 Token to the project, 80 Token will be available for the project in total. Each group member will receive 0.4*80 = 32 Token from the project.
- Assuming nobody will contribute to the project (0 Token), then nobody will receive earnings from the project since every group member decided to keep 20 Token for him- or herself.
- Assuming you contribute 5 Token to the project and each of the other group members contributes 10 Token, then you will get 0.4*(5+10+10+10) = 14 Token in addition to the 15 Token (which you did not contribute). The other group members will get 14 Token from the project as well, but they will only have 10 Token left from before.

Screen stage 1: Input of your answer into the empty box and confirmation with "Continue"

Stage 2 – Assignment of reduction points

At the beginning of this stage each group member will receive 10 additional Token. These Token can be used to reduce the earnings of the other members of the group. You can do this by assigning reduction points. At the beginning of this stage you will see how much the other group members will have contributed to the project (see monitor screen below). Afterwards, you will decide whether or not you want to assign reduction points to other group members, and in case you want to do so, how many reduction points you want to assign. You have to pay 1 Token for each reduction point you want to assign. The group member's earnings will be reduced by 3 Token for every reduction point received. More specifically, you can pay 1 Token to reduce the earnings of another member by 3 Tokens. You can distribute a maximum of 10 reduction points. The other members have the same possibility as you do.



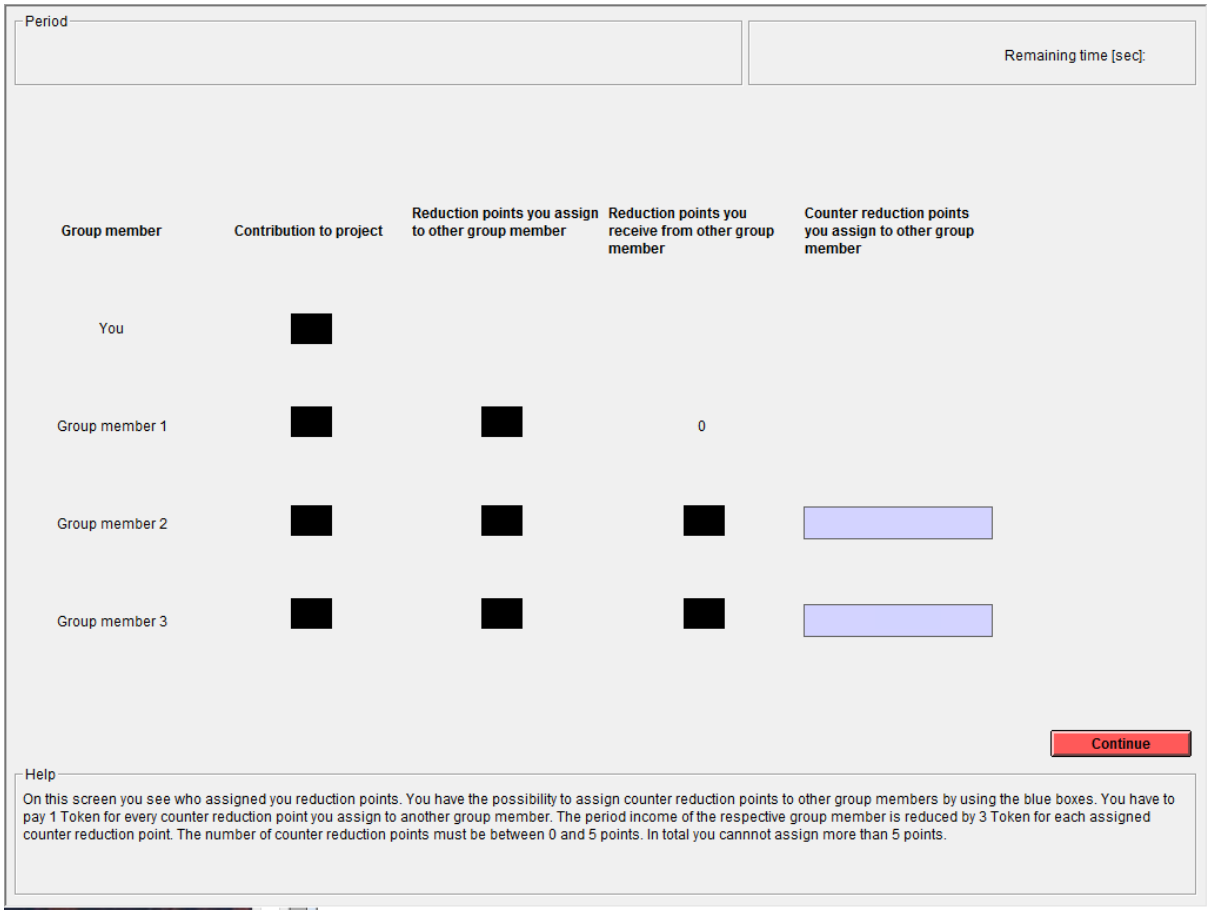
Screen stage 2: Input of the number of reduction points into the three empty boxes and confirmation with "Continue".

On the screen you will see, besides the indication of the period and the remaining time, how high the contribution of the other members was. **Your contribution** is indicated in **the first row** (labeled "you"). You will also see the contribution of the other members in the rows below. **Please be aware of the fact that the order in which the contributions of the other three group members are shown is different for each period, since the identification number for each group member will be randomly assigned every period**. More specifically, this means that the person behind the identification number 3 for example, can be a different group member from period to period. **Note the identification numbers stay the same within one period**.

Stage 3 – Assignment of counter reduction points

At the beginning of this stage each group member will receive an additional 5 Token. These Token can be used to reduce the earnings of those group members, who reduced

your own earnings in stage 2. Furthermore, you will receive information about who assigned you how many reduction points and you will also see how many Tokens were contributed by you and by the other group members. Afterwards, you can state on the respective line if you want to assign counter reduction points. If that is the case, you have to state how many counter reduction points you want to assign. The cost of assigning a counter reduction point is, as in stage 2, 1 Token per point. Each received counter reduction point will lead to a reduction of earnings of 3 Token. You can distribute a maximum of 5 counter reduction points.



Screen stage 3: Input of your answer into the empty box and confirmation with "Continue"

Overview of total earnings in one period

At the end of each period you will be informed about your earnings during that period and how it is composed (see monitor screen below). You will see the contributions to the project of all group members, the number of reduction points that you assigned to the other members, the number of received reduction points, the counter reduction points you assigned and your received counter reduction points. The formula below will show

you how the earnings are composed during one period.

Earnings in one period of a group member

=

(20 Token) – (Contribution to the project)

+

(0.4 Token * sum of all contributions to the project)

+

(10 Token) – (1 Token * number of assigned reduction points)

-

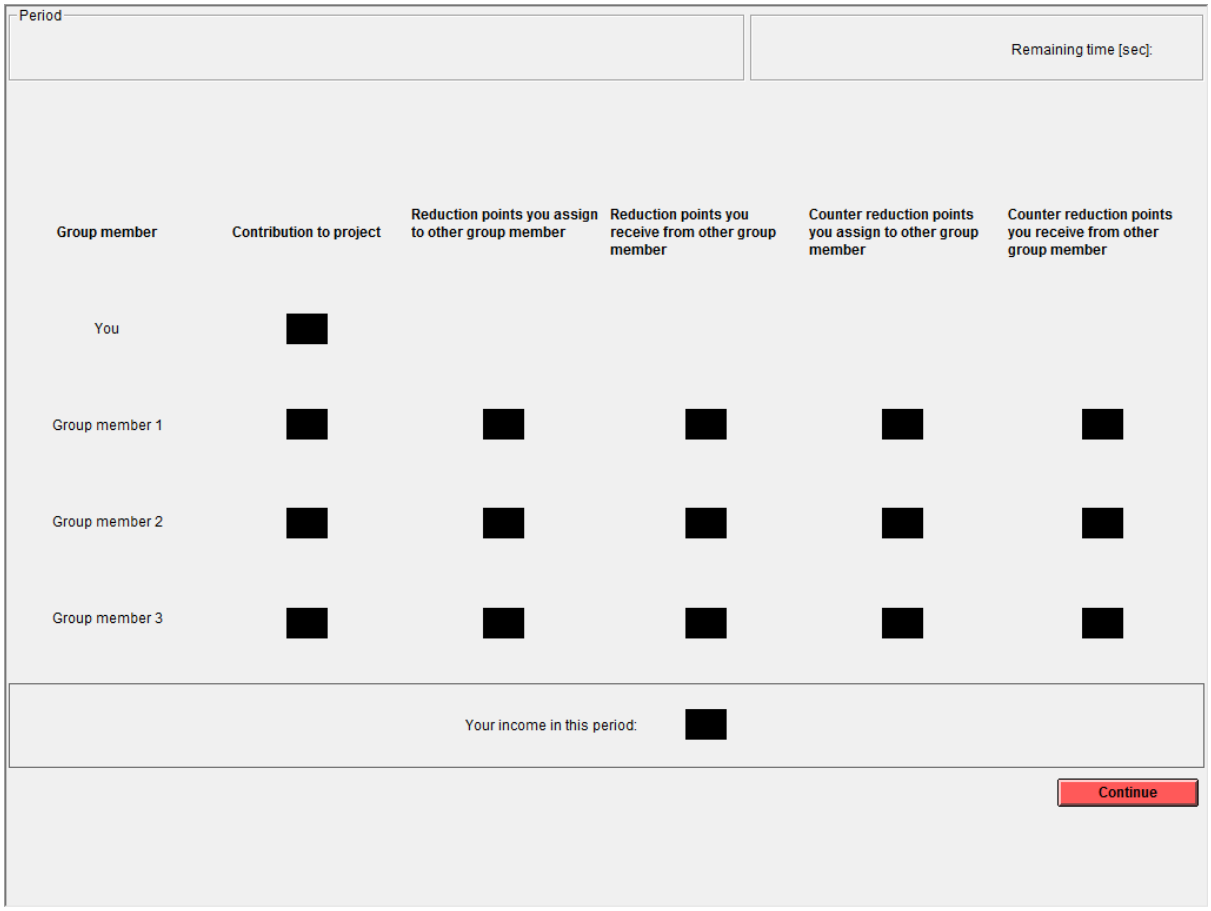
(3 Token * number of received reduction points)

+

(5 Token) - (1 Token * number of assigned counter reduction points)

-

(3 Tokens * number of received counter reduction points)



Screen at the end of each period showing an overview of the period and your period earnings. Confirmation with "Continue"

Control questions

Please answer the following questions and raise your hand as soon as you have finished.

1. Assuming nobody contributes anything (including you) to the project, and nobody assigns reduction points nor counter reduction points.

   How much are your earnings in this period? _____ Token

   How much are the earnings of the other group members? _____ Token

2. Assuming everybody contributes 20 Tokens to the project (including you). Furthermore, nobody assigns reduction points nor counter reduction points.

   How much are your earnings in this period? _____ Token

How much are the earnings of the other group members? _____ Token

3. Assuming the other three group members contribute in total 30 Token to the project. Furthermore, nobody assigns reduction points nor counter reduction points.

   a) How much are your earnings in this period if you contribute 0 Token (additionally to the 30 Token of the other members) to the project? _____ Token

   b) How much are your earnings in this period if you contribute 15 Tokens (additionally to the 30 Token of the other members) to the project?
   _____ Token

4. Assuming you contribute 8 Token to the project. Furthermore, nobody assigns reduction points nor counter reduction points.

   a) How much are your earnings if the other group members contribute in total 7 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

   b) How much are your earnings if the other group members contribute in total 22 Token to the project (in addition to your contribution of 8 Token)?
   _____ Token

5. Assuming you assign 5 reduction points to another group member.

   a) How much does this decrease the earnings of the other group member?
   _____ Token

   b) Assume another member assigns 2 counter reduction points to you. How much does this decrease your earnings?_____ Token

6. Can you assign counter reduction points during stage 3 to another member if you did not receive any reduction points from that member during stage 2?

   _____ YES _____ NO

7. Is the person with the identified as "group member 1" necessarily the same in each period?

   _____ YES _____ NO

Instructions sessions noNF–noNF Part 2

Second part of the study

Please read through the following instructions carefully. If you have any questions, please raise your hand and we will immediately send an employee to your assigned place.

Procedure of the second part

The second part of the study consists again of 15 periods. Every period consists of the **same stages** as in the first part. You will form a group together with the same three participants as in the first part. The composition of the group will therefore **not** change.

Your total income will be the sum of your earnings in the first and second part.

3 Instructions Chapter V

Instructions Dissuasion Treatment

Welcome

Thank you for participating in this study. For completion of this study you will be paid a participation fee of CHF 10. You can also earn additional money. How much money you earn (in addition to the CHF 10) depends on your decisions and chance.

Study Standards

This study is conducted by researchers at the University of Zurich. At the beginning of the session we took a picture of you. This picture will be shown to the other participants at some point during today's session. The details are outlined below. Your picture is not shown to anyone except for the participants in this session. It will be deleted immediately after the session; no copies will be retained.

All data gathered during this study will be completely anonymized. Neither the researchers nor anyone else will be able to link your behavior (e.g. answers, decisions) to any personally identifiable information (e.g. your name).

In line with the scientific standards of this lab, we will not lie to you at any time during this session.

Rules

If, at any time during the session, you have a question, please do not hesitate to ask us for help. To do so, please raise your hand, and wait for a member of staff to come to your assistance.

During the entire session, please do not talk or otherwise communicate with other participants. Please turn your mobile phone completely off. If you violate these rules, we will have to exclude you from the session and you will not be paid.

Addiotional Earnings

In addition to the participation fee, you can earn money by answering understanding questions, completing a job, and answering the survey at the end.

On your desk, there is a handout with the understanding questions. These questions are designed to help you to check whether you fully understand the instructions. You will be paid an additional CHF 10 for answering all questions (therefore, in today's study you will earn at least CHF 20).

Make sure you fully understand everything, this will allow you to earn additional money. Please do not hesitate to ask for help by raising your hand.

Next, we describe the job. It is your choice whether or not to do the job.

The Job

You have the possibility to earn money by choosing to do a job related to smoking tobacco. The job involves wrapping three cigarettes in gift wrap paper and placing them into a gift bag. The cigarettes are manufactured by British American Tobacco and sold under its Parisienne™ brand. You find the supplies needed for this job (cigarettes, gift wrap paper, ribbons, stickers, gift bag) on your desk. Also, on your desk there is an example of a wrapped cigarette.

You will be paid for doing this job. You can decide whether to do the job or not. That is, you have the choice between wrapping the three cigarettes and earning additional money, or not doing the job and not earning any additional money.

All gift bags prepared today will be distributed to young adults. These young adults participate in a short event related to tobacco. During this event, each young adult will receive one gift bag. You prepare one of these gift bags. Thus, one adult will receive your gift bag. He or she will *receive three free cigarettes* if you do the job (wrap cigarettes, place them in bag). He or she will receive a gift bag without any cigarettes if you do not do the job. Therefore, depending on your decision this young adult will or will not be exposed to this British American Tobacco product.

Procedure

1. Video: Once all participants have read the instructions, and have answered all understanding questions, the computer program will be started. The computer first shows you a video. The video is called "The Truth About Tobacco: How Much is a Life Worth?" and is produced by the American Cancer Society.

2. Your decision: After you have seen the video, you will make your decision on whether you accept or decline to do the job of wrapping 3 Parisienne™ cigarettes manufactured by British American Tobacco in gift wrap paper. We will explain in a moment what your exact choices will be. If you decline to do the job, you will not have to wrap any cigarettes, and you will not earn any additional money.

3. Survey: After all participants made their decisions, you will complete a short survey.

4. Job: Afterwards, participants will be asked to do their jobs according to their choices. Do not start working on the job before you are told to do so. Once all participants are done, we will collect all gift bags. If you accepted to do the job, we will check that you have carefully completed the job according to the instructions. If you refuse to complete the job when you in fact stated to accept it, you will not receive any payment for this study (you will not even receive your participation fee). Note that you will always have the option to decline to do the job.

5. Decisions of all participants displayed: Next, your decision whether or not to do the job will be shown to all other participants in the session, together with your picture. We will explain the details in a moment. All other participants are exactly in the same situation as you are, that is, they receive the same instructions, watch the same video, answer the same questions, face the same choices, and their choices will be shown to all participants (including you).

6. Payment: At the end of the session, all participants receive their payment in private. We ask for your understanding that you cannot leave early if you have finished fast for any reason (e.g., because you are fast at wrapping cigarettes or declined to wrap cigarettes).

Our Choices

In the following, we explain how you make your choices regarding accepting or declining to do the job.

At the end of the session, the computer will randomly select a wage between 0 and 25 CHF ("drawn wage"), in increments of 1 CHF. The drawn wage will be the same for every participant in today's session and will be the actual wage that you receive for doing the job. However, you will not know the drawn wage until the end of the session. Instead, you have to specify for each possible wage whether you accept or decline the job at this wage. At the end of the session, you will learn the drawn wage, and you will have to implement your choice that corresponds to this wage.

- If you have accepted to do the job at the drawn wage, you will have to do the job and receive the drawn wage.
- If you have declined to do the job at the drawn wage, you will not do the job, and you will not receive the drawn wage.

Recall that you will receive at least the participation fee of CHF 10 plus the CHF 10 from answering the understanding questions.

You will make your decisions on a screen that looks like the picture on page 4. Please take a moment to look at page 4, and read the following explanations:

Box "I decline to do the job for any wage"

If you tick this box, you do not have to make any other decisions in the table. Simply click OK to confirm and proceed. If you choose to tick this box, you will not wrap any cigarettes, and you will not earn any wage.

Table

Each row in the table corresponds to one possible wage. The wages are indicated in the middle column of the table in bold. You will choose, for each possible wage, whether to wrap the cigarettes or not. You will indicate your choice in each row, by either clicking the box on the left if you choose to accept the job at that wage, or the box on the right if you choose to decline the job at that wage.

At the end of the session the computer will randomly select one row in the table (drawn wage). This drawn wage is the same for everyone.

- If in that row, you indicated that you want to do the job (box on the left) at that wage, then you will do the job and earn the wage in that row.
- If in that row, you indicated that you do not want to do the job at that wage, then you will not do the job, and will not earn any additional wage.

The final payoff that you receive from accepting the job at a given wage (if this wage is randomly selected to be the drawn wage) is shown in the left column and the final payoff from declining the job at a given wage is shown in the right column.

Each row, and thus each wage, has the same probability to be drawn. Note which wage is drawn at the end of the session for you and all other participants does not depend on your choices.

The Choice Screen

□ I decline to do the job for any wage (Final Payoff = CHF 20)

Final Payoff (CHF 20 + wage)	I **accept** to do the job for this wage	**Wage (one row randomly selected as drawn wage)**	I **decline** to do the job for this wage	Final Payoff (CHF 20)
CHF 20	○	**CHF 0**	○	CHF 20
CHF 21	○	**CHF 1**	○	CHF 20
CHF 22	○	**CHF 2**	○	CHF 20
CHF 23	○	**CHF 3**	○	CHF 20
CHF 24	○	**CHF 4**	○	CHF 20
CHF 25	○	**CHF 5**	○	CHF 20
CHF 26	○	**CHF 6**	○	CHF 20
CHF 27	○	**CHF 7**	○	CHF 20
CHF 28	○	**CHF 8**	○	CHF 20
CHF 29	○	**CHF 9**	○	CHF 20
CHF 30	○	**CHF 10**	○	CHF 20
CHF 31	○	**CHF 11**	○	CHF 20
CHF 32	○	**CHF 12**	○	CHF 20
CHF 33	○	**CHF 13**	○	CHF 20
CHF 34	○	**CHF 14**	○	CHF 20
CHF 35	○	**CHF 15**	○	CHF 20
CHF 36	○	**CHF 16**	○	CHF 20
CHF 37	○	**CHF 17**	○	CHF 20
CHF 38	○	**CHF 18**	○	CHF 20
CHF 39	○	**CHF 19**	○	CHF 20
CHF 40	○	**CHF 20**	○	CHF 20
CHF 41	○	**CHF 21**	○	CHF 20
CHF 42	○	**CHF 22**	○	CHF 20
CHF 43	○	**CHF 23**	○	CHF 20
CHF 44	○	**CHF 24**	○	CHF 20
CHF 45	○	**CHF 25**	○	CHF 20

Display of your decision

At the end of the session the computer randomly selects a wage ("drawn wage') for all participants. Then your picture and your choice (for the drawn wage) will be shown to all participants in today's session. Below you can see what exactly will be shown, there are two possibilities:

If you **accepted** to wrap the cigarettes at the drawn wage:

If you **declined** to wrap the cigarettes at the drawn wage:

In the same manner as the other participants see your choice, you will see the picture of each of the other participants, linked to his or her choice at the drawn wage.

If you have any questions about the instructions, please raise your hand to get assistance. Otherwise, please turn to the handout with the understanding questions.

Understanding Questions Dissuasion

Please answer the following questions. Once you are finished, please raise your hand and wait for a member of staff to come to you.

1. Is it possible to decline to do the job for every possible wage?

_____ YES _____ NO

2. "The gift bags, including any wrapped cigarettes, will be distributed to young adults as a present"

   _____ TRUE _____ FALSE

3. "The video you are going to watch is produced by the American Cancer Society."

   _____ TRUE _____ FALSE

4. "Each other participant receives exactly the same instructions, watches the same video of the American Cancer Society, answers the same understanding questions and faces the same choices as you do."

   _____ TRUE _____ FALSE

5. "Each other participant receives exactly the same instructions, watches the same video of the American Cancer Society, answers the same understanding questions and faces the same choices as you do."

   _____ TRUE _____ FALSE

6. "At the end of the session, the computer will randomly select the wage ("drawn wage"). The drawn wage will be the same for every participant in today's session."

   _____ TRUE _____ FALSE

7. What will the other participants see about you and your choices in today's experiment? (you can choose multiple options)

   _____ Your name
   _____ Your picture
   _____ Your decision whether or not you accepted to wrap the cigarettes at each possible wage
   _____ Your decision whether or not you accepted to wrap the cigarettes at the "drawn wage"

8. "All participants leave the laboratory at the same time, that is, after all participants who accepted the job for their drawn wage are done with wrapping the cigarettes in gift wrap paper, and everyone has completed the survey"

_____ TRUE _____ FALSE

9. "For this question consider the table below. Here is an example of possible choices. *This example is purely hypothetical, it is not based on actual choices of a participant.*"

_____ TRUE _____ FALSE

□ I decline to do the job for any wage (Final Payoff = CHF 20)

Final Payoff (CHF 20 + wage)	I **accept** to do the job for this wage	**Wage (one row randomly selected as drawn wage)**	I **decline** to do the job for this wage	Final Payoff (CHF 20)
CHF 20	○	**CHF 0**	●	CHF 20
CHF 21	○	**CHF 1**	●	CHF 20
CHF 22	○	**CHF 2**	●	CHF 20
CHF 23	○	**CHF 3**	●	CHF 20
CHF 24	○	**CHF 4**	●	CHF 20
CHF 25	○	**CHF 5**	●	CHF 20
CHF 26	○	**CHF 6**	●	CHF 20
CHF 27	○	**CHF 7**	●	CHF 20
CHF 28	○	**CHF 8**	●	CHF 20
CHF 29	○	**CHF 9**	●	CHF 20
CHF 30	○	**CHF 10**	●	CHF 20
CHF 31	●	**CHF 11**	○	CHF 20
CHF 32	●	**CHF 12**	○	CHF 20
CHF 33	●	**CHF 13**	○	CHF 20
CHF 34	●	**CHF 14**	○	CHF 20
CHF 35	●	**CHF 15**	○	CHF 20
CHF 36	●	**CHF 16**	○	CHF 20
CHF 37	●	**CHF 17**	○	CHF 20
CHF 38	●	**CHF 18**	○	CHF 20
CHF 39	●	**CHF 19**	○	CHF 20
CHF 40	●	**CHF 20**	○	CHF 20
CHF 41	●	**CHF 21**	○	CHF 20
CHF 42	●	**CHF 22**	○	CHF 20
CHF 43	●	**CHF 23**	○	CHF 20
CHF 44	●	**CHF 24**	○	CHF 20
CHF 45	●	**CHF 25**	○	CHF 20

a) "If the computer randomly selects the row with a wage of CHF 11, the participant does the job." _____ TRUE _____ FALSE

b) Suppose the computer randomly selects the row with a wage of CHF 11. Then,

what is the final payoff? CHF _____

c) Suppose the computer randomly selects the row with a wage of CHF 10. Then, what is the final payoff? CHF _____

Instructions Neutral Treatment

Welcome

Thank you for participating in this study. For completion of this study you will be paid a participation fee of CHF 10. You can also earn additional money. How much money you earn (in addition to the CHF 10) depends on your decisions and chance.

Study Standards

This study is conducted by researchers at the University of Zurich. At the beginning of the session we took a picture of you. This picture will be shown to the other participants at some point during today's session. The details are outlined below. Your picture is not shown to anyone except for the participants in this session. It will be deleted immediately after the session; no copies will be retained.

All data gathered during this study will be completely anonymized. Neither the researchers nor anyone else will be able to link your behavior (e.g. answers, decisions) to any personally identifiable information (e.g. your name).

In line with the scientific standards of this lab, we will not lie to you at any time during this session.

Rules

If, at any time during the session, you have a question, please do not hesitate to ask us for help. To do so, please raise your hand, and wait for a member of staff to come to your assistance.

During the entire session, please do not talk or otherwise communicate with other participants. Please turn your mobile phone completely off. If you violate these rules, we will have to exclude you from the session and you will not be paid.

Addiotional Earnings

In addition to the participation fee, you can earn money by answering understanding questions, completing a job, and answering the survey at the end.

On your desk, there is a handout with the understanding questions. These questions are designed to help you to check whether you fully understand the instructions. You will be paid an additional CHF 10 for answering all questions (therefore, in today's study you will earn at least CHF 20).

Make sure you fully understand everything, this will allow you to earn additional money. Please do not hesitate to ask for help by raising your hand.

Next, we describe the job. It is your choice whether or not to do the job.

The Job

You have the possibility to earn money by choosing to do a job related to smoking tobacco. The job involves wrapping three cigarettes in gift wrap paper and placing them into a gift bag. The cigarettes are manufactured by British American Tobacco and sold under its Parisienne™ brand. You find the supplies needed for this job (cigarettes, gift wrap paper, ribbons, stickers, gift bag) on your desk. Also, on your desk there is an example of a wrapped cigarette.

You will be paid for doing this job. You can decide whether to do the job or not. That is, you have the choice between wrapping the three cigarettes and earning additional money, or not doing the job and not earning any additional money.

All gift bags prepared today will be distributed to young adults. These young adults participate in a short event related to tobacco. During this event, each young adult will receive one gift bag. You prepare one of these gift bags. Thus, one adult will receive your gift bag. He or she will *receive three free cigarettes* if you do the job (wrap cigarettes, place them in bag). He or she will receive a gift bag without any cigarettes if you do not do the job. Therefore, depending on your decision this young adult will or will not be exposed to this British American Tobacco product.

Procedure

1. Once all participants have read the instructions, and have answered all understanding questions, the computer program will be started. The computer first shows you a video. The video is called "Scandinavia: Landscapes of Sweden" and is produced by an independent content producer for non-commercial purposes.

2. Your decision: After you have seen the video, you will make your decision on whether you accept or decline to do the job of wrapping 3 Parisienne™ cigarettes manufactured by British American Tobacco in gift wrap paper. We will explain in a moment what your exact choices will be. If you decline to do the job, you will not have to wrap any cigarettes, and you will not earn any additional money.

3. Survey: After all participants made their decisions, you will complete a short survey.

4. Job: Afterwards, participants will be asked to do their jobs according to their choices. Do not start working on the job before you are told to do so. Once all participants are done, we will collect all gift bags. If you accepted to do the job, we will check that you have carefully completed the job according to the instructions. If you refuse to complete the job when you in fact stated to accept it, you will not receive any payment for this study (you will not even receive your participation fee). Note that you will always have the option to decline to do the job.

5. Decisions of all participants displayed: Next, your decision whether or not to do the job will be shown to all other participants in the session, together with your picture. We will explain the details in a moment. All other participants are exactly in the same situation as you are, that is, they receive the same instructions, watch the same video, answer the same questions, face the same choices, and their choices will be shown to all participants (including you).

6. Payment: At the end of the session, all participants receive their payment in private. We ask for your understanding that you cannot leave early if you have finished fast for any reason (e.g., because you are fast at wrapping cigarettes or declined to wrap cigarettes).

Our Choices

In the following, we explain how you make your choices regarding accepting or declining to do the job.

At the end of the session, the computer will randomly select a wage between 0 and 25 CHF ("drawn wage"), in increments of 1 CHF. The drawn wage will be the same for every participant in today's session and will be the actual wage that you receive for doing the job. However, you will not know the drawn wage until the end of the session. Instead, you have to specify for each possible wage whether you accept or decline the job at this wage. At the end of the session, you will learn the drawn wage, and you will have to implement your choice that corresponds to this wage.

- If you have accepted to do the job at the drawn wage, you will have to do the job and receive the drawn wage.
- If you have declined to do the job at the drawn wage, you will not do the job, and you will not receive the drawn wage.

Recall that you will receive at least the participation fee of CHF 10 plus the CHF 10 from answering the understanding questions.

You will make your decisions on a screen that looks like the picture on page 4. Please take a moment to look at page 4, and read the following explanations:

Box "I decline to do the job for any wage"

If you tick this box, you do not have to make any other decisions in the table. Simply click OK to confirm and proceed. If you choose to tick this box, you will not wrap any cigarettes, and you will not earn any wage.

Table

Each row in the table corresponds to one possible wage. The wages are indicated in the middle column of the table in bold. You will choose, for each possible wage, whether to wrap the cigarettes or not. You will indicate your choice in each row, by either clicking the box on the left if you choose to accept the job at that wage, or the box on the right if you choose to decline the job at that wage.

At the end of the session the computer will randomly select one row in the table (drawn wage). This drawn wage is the same for everyone.

- If in that row, you indicated that you want to do the job (box on the left) at that wage, then you will do the job and earn the wage in that row.
- If in that row, you indicated that you do not want to do the job at that wage, then you will not do the job, and will not earn any additional wage.

The final payoff that you receive from accepting the job at a given wage (if this wage is randomly selected to be the drawn wage) is shown in the left column and the final payoff from declining the job at a given wage is shown in the right column.

Each row, and thus each wage, has the same probability to be drawn. Note which wage is drawn at the end of the session for you and all other participants does not depend on your choices.

The Choice Screen

□ I decline to do the job for any wage (Final Payoff = CHF 20)

Final Payoff (CHF 20 + wage)	I **accept** to do the job for this wage	**Wage (one row randomly selected as drawn wage)**	I **decline** to do the job for this wage	Final Payoff (CHF 20)
CHF 20	○	**CHF 0**	○	CHF 20
CHF 21	○	**CHF 1**	○	CHF 20
CHF 22	○	**CHF 2**	○	CHF 20
CHF 23	○	**CHF 3**	○	CHF 20
CHF 24	○	**CHF 4**	○	CHF 20
CHF 25	○	**CHF 5**	○	CHF 20
CHF 26	○	**CHF 6**	○	CHF 20
CHF 27	○	**CHF 7**	○	CHF 20
CHF 28	○	**CHF 8**	○	CHF 20
CHF 29	○	**CHF 9**	○	CHF 20
CHF 30	○	**CHF 10**	○	CHF 20
CHF 31	○	**CHF 11**	○	CHF 20
CHF 32	○	**CHF 12**	○	CHF 20
CHF 33	○	**CHF 13**	○	CHF 20
CHF 34	○	**CHF 14**	○	CHF 20
CHF 35	○	**CHF 15**	○	CHF 20
CHF 36	○	**CHF 16**	○	CHF 20
CHF 37	○	**CHF 17**	○	CHF 20
CHF 38	○	**CHF 18**	○	CHF 20
CHF 39	○	**CHF 19**	○	CHF 20
CHF 40	○	**CHF 20**	○	CHF 20
CHF 41	○	**CHF 21**	○	CHF 20
CHF 42	○	**CHF 22**	○	CHF 20
CHF 43	○	**CHF 23**	○	CHF 20
CHF 44	○	**CHF 24**	○	CHF 20
CHF 45	○	**CHF 25**	○	CHF 20

Display of your decision

At the end of the session the computer randomly selects a wage ("drawn wage') for all participants. Then your picture and your choice (for the drawn wage) will be shown to all participants in today's session. Below you can see what exactly will be shown, there are two possibilities:

If you **accepted** to wrap the cigarettes at the drawn wage:

If you **declined** to wrap the cigarettes at the drawn wage:

In the same manner as the other participants see your choice, you will see the picture of each of the other participants, linked to his or her choice at the drawn wage.

If you have any questions about the instructions, please raise your hand to get assistance. Otherwise, please turn to the handout with the understanding questions.

Understanding Questions Neutral

Please answer the following questions. Once you are finished, please raise your hand and wait for a member of staff to come to you.

1. Is it possible to decline to do the job for every possible wage?

_____ YES _____ NO

2. "The gift bags, including any wrapped cigarettes, will be distributed to young adults as a present"

   _____ TRUE _____ FALSE

3. "The video you are going to watch is called *Scandinavia: Landscapes of Sweden*."

   _____ TRUE _____ FALSE

4. "Each other participant receives exactly the same instructions, watches the same video of the American Cancer Society, answers the same understanding questions and faces the same choices as you do."

   _____ TRUE _____ FALSE

5. "Each other participant receives exactly the same instructions, watches the same video of the American Cancer Society, answers the same understanding questions and faces the same choices as you do."

   _____ TRUE _____ FALSE

6. "At the end of the session, the computer will randomly select the wage ("drawn wage"). The drawn wage will be the same for every participant in today's session."

   _____ TRUE _____ FALSE

7. What will the other participants see about you and your choices in today's experiment? (you can choose multiple options)

   _____ Your name
   _____ Your picture
   _____ Your decision whether or not you accepted to wrap the cigarettes at each possible wage
   _____ Your decision whether or not you accepted to wrap the cigarettes at the "drawn wage"

8. "All participants leave the laboratory at the same time, that is, after all participants who accepted the job for their drawn wage are done with wrapping the cigarettes in gift wrap paper, and everyone has completed the survey"

_____ TRUE _____ FALSE

9. "For this question consider the table below. Here is an example of possible choices. *This example is purely hypothetical, it is not based on actual choices of a participant.*"

_____ TRUE _____ FALSE

□ I decline to do the job for any wage (Final Payoff = CHF 20)

Final Payoff (CHF 20 + wage)	I **accept** to do the job for this wage	**Wage (one row randomly selected as drawn wage)**	I **decline** to do the job for this wage	Final Payoff (CHF 20)
CHF 20	○	**CHF 0**	●	CHF 20
CHF 21	○	**CHF 1**	●	CHF 20
CHF 22	○	**CHF 2**	●	CHF 20
CHF 23	○	**CHF 3**	●	CHF 20
CHF 24	○	**CHF 4**	●	CHF 20
CHF 25	○	**CHF 5**	●	CHF 20
CHF 26	○	**CHF 6**	●	CHF 20
CHF 27	○	**CHF 7**	●	CHF 20
CHF 28	○	**CHF 8**	●	CHF 20
CHF 29	○	**CHF 9**	●	CHF 20
CHF 30	○	**CHF 10**	●	CHF 20
CHF 31	●	**CHF 11**	○	CHF 20
CHF 32	●	**CHF 12**	○	CHF 20
CHF 33	●	**CHF 13**	○	CHF 20
CHF 34	●	**CHF 14**	○	CHF 20
CHF 35	●	**CHF 15**	○	CHF 20
CHF 36	●	**CHF 16**	○	CHF 20
CHF 37	●	**CHF 17**	○	CHF 20
CHF 38	●	**CHF 18**	○	CHF 20
CHF 39	●	**CHF 19**	○	CHF 20
CHF 40	●	**CHF 20**	○	CHF 20
CHF 41	●	**CHF 21**	○	CHF 20
CHF 42	●	**CHF 22**	○	CHF 20
CHF 43	●	**CHF 23**	○	CHF 20
CHF 44	●	**CHF 24**	○	CHF 20
CHF 45	●	**CHF 25**	○	CHF 20

a) "If the computer randomly selects the row with a wage of CHF 11, the participant does the job." _____ TRUE _____ FALSE

b) Suppose the computer randomly selects the row with a wage of CHF 11. Then,

what is the final payoff? CHF _____

c) Suppose the computer randomly selects the row with a wage of CHF 10. Then, what is the final payoff? CHF _____

Instructions Persuasion Treatment

Welcome

Thank you for participating in this study. For completion of this study you will be paid a participation fee of CHF 10. You can also earn additional money. How much money you earn (in addition to the CHF 10) depends on your decisions and chance.

Study Standards

This study is conducted by researchers at the University of Zurich. At the beginning of the session we took a picture of you. This picture will be shown to the other participants at some point during today's session. The details are outlined below. Your picture is not shown to anyone except for the participants in this session. It will be deleted immediately after the session; no copies will be retained.

All data gathered during this study will be completely anonymized. Neither the researchers nor anyone else will be able to link your behavior (e.g. answers, decisions) to any personally identifiable information (e.g. your name).

In line with the scientific standards of this lab, we will not lie to you at any time during this session.

Rules

If, at any time during the session, you have a question, please do not hesitate to ask us for help. To do so, please raise your hand, and wait for a member of staff to come to your assistance.

During the entire session, please do not talk or otherwise communicate with other participants. Please turn your mobile phone completely off. If you violate these rules, we will have to exclude you from the session and you will not be paid.

Addiotional Earnings

In addition to the participation fee, you can earn money by answering understanding questions, completing a job, and answering the survey at the end.

On your desk, there is a handout with the understanding questions. These questions are designed to help you to check whether you fully understand the instructions. You will be paid an additional CHF 10 for answering all questions (therefore, in today's study you will earn at least CHF 20).

Make sure you fully understand everything, this will allow you to earn additional money. Please do not hesitate to ask for help by raising your hand.

Next, we describe the job. It is your choice whether or not to do the job.

The Job

You have the possibility to earn money by choosing to do a job related to smoking tobacco. The job involves wrapping three cigarettes in gift wrap paper and placing them into a gift bag. The cigarettes are manufactured by British American Tobacco and sold under its Parisienne™ brand. You find the supplies needed for this job (cigarettes, gift wrap paper, ribbons, stickers, gift bag) on your desk. Also, on your desk there is an example of a wrapped cigarette.

You will be paid for doing this job. You can decide whether to do the job or not. That is, you have the choice between wrapping the three cigarettes and earning additional money, or not doing the job and not earning any additional money.

All gift bags prepared today will be distributed to young adults. These young adults participate in a short event related to tobacco. During this event, each young adult will receive one gift bag. You prepare one of these gift bags. Thus, one adult will receive your gift bag. He or she will *receive three free cigarettes* if you do the job (wrap cigarettes, place them in bag). He or she will receive a gift bag without any cigarettes if you do not do the job. Therefore, depending on your decision this young adult will or will not be exposed to this British American Tobacco product.

Procedure

1. Once all participants have read the instructions, and have answered all understanding questions, the computer program will be started. The computer first shows you a video. The video is the official company video of British American Tobacco, the manufacturer of Parisienne™ cigarettes. British American Tobacco features this video on its homepage and their official corporate channel on YouTube.

2. Your decision: After you have seen the video, you will make your decision on whether you accept or decline to do the job of wrapping 3 Parisienne™ cigarettes manufactured by British American Tobacco in gift wrap paper. We will explain in a moment what your exact choices will be. If you decline to do the job, you will not have to wrap any cigarettes, and you will not earn any additional money.

3. Survey: After all participants made their decisions, you will complete a short survey.

4. Job: Afterwards, participants will be asked to do their jobs according to their choices. Do not start working on the job before you are told to do so. Once all participants are done, we will collect all gift bags. If you accepted to do the job, we will check that you have carefully completed the job according to the instructions. If you refuse to complete the job when you in fact stated to accept it, you will not receive any payment for this study (you will not even receive your participation fee). Note that you will always have the option to decline to do the job.

5. Decisions of all participants displayed: Next, your decision whether or not to do the job will be shown to all other participants in the session, together with your picture. We will explain the details in a moment. All other participants are exactly in the same situation as you are, that is, they receive the same instructions, watch the same video, answer the same questions, face the same choices, and their choices will be shown to all participants (including you).

6. Payment: At the end of the session, all participants receive their payment in private. We ask for your understanding that you cannot leave early if you have finished fast for any reason (e.g., because you are fast at wrapping cigarettes or declined to wrap cigarettes).

Our Choices

In the following, we explain how you make your choices regarding accepting or declining to do the job.

At the end of the session, the computer will randomly select a wage between 0 and 25 CHF ("drawn wage"), in increments of 1 CHF. The drawn wage will be the same for every participant in today's session and will be the actual wage that you receive for doing the job. However, you will not know the drawn wage until the end of the session. Instead, you have to specify for each possible wage whether you accept or decline the job at this wage. At the end of the session, you will learn the drawn wage, and you will have to implement your choice that corresponds to this wage.

- If you have accepted to do the job at the drawn wage, you will have to do the job and receive the drawn wage.
- If you have declined to do the job at the drawn wage, you will not do the job, and you will not receive the drawn wage.

Recall that you will receive at least the participation fee of CHF 10 plus the CHF 10 from answering the understanding questions.

You will make your decisions on a screen that looks like the picture on page 4. Please take a moment to look at page 4, and read the following explanations:

Box "I decline to do the job for any wage"

If you tick this box, you do not have to make any other decisions in the table. Simply click OK to confirm and proceed. If you choose to tick this box, you will not wrap any cigarettes, and you will not earn any wage.

Table

Each row in the table corresponds to one possible wage. The wages are indicated in the middle column of the table in bold. You will choose, for each possible wage, whether to wrap the cigarettes or not. You will indicate your choice in each row, by either clicking the box on the left if you choose to accept the job at that wage, or the box on the right if you choose to decline the job at that wage.

At the end of the session the computer will randomly select one row in the table (drawn wage). This drawn wage is the same for everyone.

- If in that row, you indicated that you want to do the job (box on the left) at that wage, then you will do the job and earn the wage in that row.
- If in that row, you indicated that you do not want to do the job at that wage, then you will not do the job, and will not earn any additional wage.

The final payoff that you receive from accepting the job at a given wage (if this wage is randomly selected to be the drawn wage) is shown in the left column and the final payoff from declining the job at a given wage is shown in the right column.

Each row, and thus each wage, has the same probability to be drawn. Note which wage is drawn at the end of the session for you and all other participants does not depend on your choices.

The Choice Screen

□ I decline to do the job for any wage (Final Payoff = CHF 20)

Final Payoff (CHF 20 + wage)	I **accept** to do the job for this wage	**Wage (one row randomly selected as drawn wage)**	I **decline** to do the job for this wage	Final Payoff (CHF 20)
CHF 20	○	**CHF 0**	○	CHF 20
CHF 21	○	**CHF 1**	○	CHF 20
CHF 22	○	**CHF 2**	○	CHF 20
CHF 23	○	**CHF 3**	○	CHF 20
CHF 24	○	**CHF 4**	○	CHF 20
CHF 25	○	**CHF 5**	○	CHF 20
CHF 26	○	**CHF 6**	○	CHF 20
CHF 27	○	**CHF 7**	○	CHF 20
CHF 28	○	**CHF 8**	○	CHF 20
CHF 29	○	**CHF 9**	○	CHF 20
CHF 30	○	**CHF 10**	○	CHF 20
CHF 31	○	**CHF 11**	○	CHF 20
CHF 32	○	**CHF 12**	○	CHF 20
CHF 33	○	**CHF 13**	○	CHF 20
CHF 34	○	**CHF 14**	○	CHF 20
CHF 35	○	**CHF 15**	○	CHF 20
CHF 36	○	**CHF 16**	○	CHF 20
CHF 37	○	**CHF 17**	○	CHF 20
CHF 38	○	**CHF 18**	○	CHF 20
CHF 39	○	**CHF 19**	○	CHF 20
CHF 40	○	**CHF 20**	○	CHF 20
CHF 41	○	**CHF 21**	○	CHF 20
CHF 42	○	**CHF 22**	○	CHF 20
CHF 43	○	**CHF 23**	○	CHF 20
CHF 44	○	**CHF 24**	○	CHF 20
CHF 45	○	**CHF 25**	○	CHF 20

Display of your decision

At the end of the session the computer randomly selects a wage ("drawn wage') for all participants. Then your picture and your choice (for the drawn wage) will be shown to all participants in today's session. Below you can see what exactly will be shown, there are two possibilities:

If you **accepted** to wrap the cigarettes at the drawn wage:

If you **declined** to wrap the cigarettes at the drawn wage:

In the same manner as the other participants see your choice, you will see the picture of each of the other participants, linked to his or her choice at the drawn wage.

If you have any questions about the instructions, please raise your hand to get assistance. Otherwise, please turn to the handout with the understanding questions.

Understanding Questions Persuasion

Please answer the following questions. Once you are finished, please raise your hand and wait for a member of staff to come to you.

1. Is it possible to decline to do the job for every possible wage?

_____ YES _____ NO

2. "The gift bags, including any wrapped cigarettes, will be distributed to young adults as a present"

   _____ TRUE _____ FALSE

3. "The video you are going to watch is the official company video of British American Tobacco, the manufacturer of Parisienne™ cigarettes."

   _____ TRUE _____ FALSE

4. "Each other participant receives exactly the same instructions, watches the same video of British American Tobacco (the official company video), answers the same understanding questions and faces the same choices as you do."

   _____ TRUE _____ FALSE

5. "Each other participant receives exactly the same instructions, watches the same video of the American Cancer Society, answers the same understanding questions and faces the same choices as you do."

   _____ TRUE _____ FALSE

6. "At the end of the session, the computer will randomly select the wage ("drawn wage"). The drawn wage will be the same for every participant in today's session."

   _____ TRUE _____ FALSE

7. What will the other participants see about you and your choices in today's experiment? (you can choose multiple options)

   _____ Your name
   _____ Your picture
   _____ Your decision whether or not you accepted to wrap the cigarettes at each possible wage
   _____ Your decision whether or not you accepted to wrap the cigarettes at the "drawn wage"

8. "All participants leave the laboratory at the same time, that is, after all participants who accepted the job for their drawn wage are done with wrapping the cigarettes in gift wrap paper, and everyone has completed the survey"

_____ TRUE _____ FALSE

9. "For this question consider the table below. Here is an example of possible choices. *This example is purely hypothetical, it is not based on actual choices of a participant.*"

_____ TRUE _____ FALSE

□ I decline to do the job for any wage (Final Payoff = CHF 20)

Final Payoff (CHF 20 + wage)	I **accept** to do the job for this wage	**Wage (one row randomly selected as drawn wage)**	I **decline** to do the job for this wage	Final Payoff (CHF 20)
CHF 20	○	**CHF 0**	●	CHF 20
CHF 21	○	**CHF 1**	●	CHF 20
CHF 22	○	**CHF 2**	●	CHF 20
CHF 23	○	**CHF 3**	●	CHF 20
CHF 24	○	**CHF 4**	●	CHF 20
CHF 25	○	**CHF 5**	●	CHF 20
CHF 26	○	**CHF 6**	●	CHF 20
CHF 27	○	**CHF 7**	●	CHF 20
CHF 28	○	**CHF 8**	●	CHF 20
CHF 29	○	**CHF 9**	●	CHF 20
CHF 30	○	**CHF 10**	●	CHF 20
CHF 31	●	**CHF 11**	○	CHF 20
CHF 32	●	**CHF 12**	○	CHF 20
CHF 33	●	**CHF 13**	○	CHF 20
CHF 34	●	**CHF 14**	○	CHF 20
CHF 35	●	**CHF 15**	○	CHF 20
CHF 36	●	**CHF 16**	○	CHF 20
CHF 37	●	**CHF 17**	○	CHF 20
CHF 38	●	**CHF 18**	○	CHF 20
CHF 39	●	**CHF 19**	○	CHF 20
CHF 40	●	**CHF 20**	○	CHF 20
CHF 41	●	**CHF 21**	○	CHF 20
CHF 42	●	**CHF 22**	○	CHF 20
CHF 43	●	**CHF 23**	○	CHF 20
CHF 44	●	**CHF 24**	○	CHF 20
CHF 45	●	**CHF 25**	○	CHF 20

a) "If the computer randomly selects the row with a wage of CHF 11, the participant does the job." _____ TRUE _____ FALSE

b) Suppose the computer randomly selects the row with a wage of CHF 11. Then,

what is the final payoff? CHF _____

c) Suppose the computer randomly selects the row with a wage of CHF 10. Then, what is the final payoff? CHF _____

Curriculum Vitae

Ivo Schurtenberger

Date of birth 01.08.1988

Education

August 16 – October 18	**Zurich Graduate School of Economics** PhD program in Economics (Track C)
September 13 – July 16	**University of Zurich (UZH)** Master of Science in Business and Economics (Track C)
September 09 – September 12	**University of Zurich (UZH)** Bachelor of Arts in Economics and Business Administration
August 04 – December 07	**Gymnasium Muttenz** Baccalaureate

Professional Experience

August 14 – September 18	**University of Zurich (UZH)** Assistant at the Department of Economics
November 17 – March 18	**Baudacci Nigg Stenberg Attorneys at Law** Economic consultant (mandate)
January 13 – June 13	**Swiss Tropical and Public Health Institute** Economic consultant in Tajikistan
July 12 – December 12	**Zurich University of Applied Sciences** Scientific collaborator at the Institute of Natural Resource Sciences